



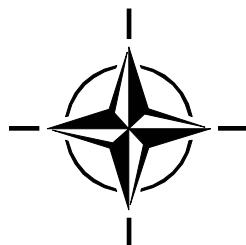
RTO TECHNICAL REPORT

TR-AVT-051

The Use of In-Service Inspection Data in the Performance Measurement of Non-Destructive Inspections

(Mise en œuvre de données résultant de
visites d'inspection en service pour
l'évaluation des performances des
visites d'inspection non destructives)

Work performed by the
RTO Applied Vehicle Technology Panel (AVT) TG-051.



Published March 2005





RTO TECHNICAL REPORT

TR-AVT-051

The Use of In-Service Inspection Data in the Performance Measurement of Non-Destructive Inspections

(Mise en œuvre de données résultant de
visites d'inspection en service pour
l'évaluation des performances des
visites d'inspection non destructives)

Work performed by the
RTO Applied Vehicle Technology Panel (AVT) TG-051.

The Research and Technology Organisation (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote co-operative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective co-ordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also co-ordinates RTO's co-operation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of co-operation.

The total spectrum of R&T activities is covered by the following 7 bodies:

- AVT Applied Vehicle Technology Panel
- HFM Human Factors and Medicine Panel
- IST Information Systems Technology Panel
- NMSG NATO Modelling and Simulation Group
- SAS Studies, Analysis and Simulation Panel
- SCI Systems Concepts and Integration Panel
- SET Sensors and Electronics Technology Panel

These bodies are made up of national representatives as well as generally recognised 'world class' scientists. They also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier co-operation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

The content of this publication has been reproduced directly from material supplied by RTO or the authors.

Published March 2005

Copyright © RTO/NATO 2005
All Rights Reserved

ISBN 92-837-1134-3

Single copies of this publication or of a part of it may be made for individual use only. The approval of the RTA Information Management Systems Branch is required for more than one copy to be made or an extract included in another publication. Requests to do so should be sent to the address on the back cover.

The Use of In-Service Inspection Data in the Performance Measurement of Non-Destructive Inspections

(RTO-TR-AVT-051)

Executive Summary

Background

Most available nondestructive inspection (NDI) reliability data results from dedicated round-robin inspection programs, wherein the same samples are inspected by disparate technicians under laboratory, or in some cases, simulated in-service conditions. These data have been frequently challenged on the basis of non-representativeness of the inspection conditions in terms of environment, access, and human factors. Analysis of in-service NDI findings can improve our understanding of the performance of NDI. This greater confidence in NDI reliability would allow more effective use of NDI for life extension.

Significant numbers of in-service inspections are occurring but at present, there is no organized process whereby these data are collected and collated for NDI reliability studies. There is undoubtedly a large amount of existing inspection data that cannot be accessed directly. The extent of this data and its usefulness to the NDI reliability program must be understood and for this reason, the Workshop “Quantification of Airframe Inspection Reliability under Field Conditions” was held in Brussels in May 1998. The processes under which this data could be collected must be defined and implemented in practical and cost effective terms. Equally, the analytical methods used to calculate NDI reliability from in-service must be defined and validated.

Summary of Findings

The AVT-051 team was multidisciplinary in that it included aircraft operators, designers, regulators, NDI specialists and statisticians. One major contribution of this work is a detailed summary of the close relationship between NDI, fracture mechanics and airworthiness including an important review of the statistical basis for many of current approaches to inspection.

With respect to the specific issue of inspection reliability from in-service inspection data, to obtain a sufficient number of cracks for a reasonable POD analysis, it would likely be necessary to pool inspection data from different sources. Accordingly, NDI maintenance records were reviewed. It was concluded that such records vary considerably in quality and fidelity. Specific recommendations for increased vigor in NDI procedure validation, calibration, application and documentation were made.

Three approaches for using in-service inspection data to characterize the capability of an inspection system were explored. Two of the approaches were directed at characterizing NDI capability in terms of the probability of detection (POD). The third approach was a direct summary of the inspection results in terms of the cumulative distribution function (CDF) of the sizes of the detected cracks.

With respect to the POD characterization of inspection capability, it was concluded that the proposed approaches to the use of in-service inspection data should not be used. In a practical maintenance scenario,

there will always be too many cracks of a detectable size that are not detected. Ignoring these “missing” misses results in a non-conservative POD characterization and the degree of non-conservatism is indeterminate. The question of the effect of errors resulting from the back calculation of crack sizes was addressed but found to be a second order effect when compared to the “missing” misses.

The CDF of detected crack sizes does provide information about the capability of the NDI system in the in-service environment. The CDF does not directly yield the reliably detectable crack size (at a given confidence level) but it gives a first estimate of this size.

Mise en œuvre de données résultant de visites d'inspection en service pour l'évaluation des performances des visites d'inspection non destructives

(RTO-TR-AVT-051)

Synthèse

Introduction

La majorité des données sur la fiabilité des visites d'inspection non destructive (NDI) sont le résultat de programmes d'inspection comparatifs, où les mêmes échantillons sont examinés par divers techniciens dans des conditions de laboratoire et, parfois, dans des conditions opérationnelles simulées. Ces données ont souvent été mises en question en raison du caractère non représentatif des conditions d'inspection du point de vue de l'environnement, de l'accès, des facteurs humains. L'analyse des conclusions des inspections NDI en service permettrait de mieux comprendre les performances des NDI. La plus grande confiance en la fiabilité des NDI qui en résulterait permettrait une meilleure utilisation de ces inspections pour le prolongement du cycle de vie.

Les visites d'inspection en service sont courantes, mais il n'existe à présent aucun processus réglementaire pour la collecte et le classement de ces données en vue de la réalisation d'études NDI. En outre, il existe aussi un volume important de données d'inspection qui ne peuvent pas être consultées directement. Or, il est essentiel de déterminer le volume et l'intérêt de ces données vis-à-vis du programme de fiabilité des NDI. C'est la raison pour laquelle l'atelier sur « La quantification de la fiabilité des visites d'inspection des cellules en conditions naturelles » a été organisé à Bruxelles en mai 1998. Les processus qui permettraient la collecte de ces données doivent être définis et mis en œuvre de façon pratique et rentable. De la même façon, les méthodes analytiques employées pour le calcul de la fiabilité des NDI en service doivent être définies et validées.

Résumé des conclusions

L'équipe AVT-051 était pluridisciplinaire, comprenant des exploitants, des concepteurs, des contrôleurs, des spécialistes en NDI et des statisticiens. L'une des contributions majeures à ces travaux a été la fourniture d'un résumé détaillé des liens étroits qui existent entre la NDI, la mécanique de la fracture et l'aptitude au vol, y compris un sommaire important des principes statistiques qui sous-tendent bon nombre des approches actuelles de l'inspection.

En ce qui concerne la question particulière de la fiabilité des visites d'inspection basée sur les données d'inspection en service, il a été considéré qu'il faudrait mettre en commun des données d'inspection obtenues de différentes sources afin d'obtenir un nombre suffisant de fissures pour permettre la réalisation d'une analyse POD dans des conditions acceptables. Par conséquent, il a été procédé à l'examen de fiches de maintenance NDI. Des variations considérables ont été constatées au niveau de la qualité et de la fidélité de ces fiches. Des recommandations particulières ont été faites concernant la nécessité de multiplier les efforts en validation des procédures NDI, en étalonnage, en applications et en documentation.

Trois approches de la mise en œuvre de données d'inspection en service pour caractériser la capacité d'un système d'inspection ont été examinées. Deux de ces approches visaient la caractérisation de la NDI en termes de probabilité de détection (POD). La troisième consistait en un résumé direct des résultats d'inspection eu égard à la fonction cumulative de distribution (CDF) des dimensions des fissures détectées.

En ce qui concerne la caractérisation POD de la capacité d'inspection, il a été conclu que les approches proposées concernant la mise en œuvre de données d'inspection en service étaient à exclure. En effet, dans un scénario de maintenance réel, il y aurait toujours trop de fissures de dimensions détectables qui ne seraient pas détectées. Ne pas tenir compte de ces « loupés » mène à des caractérisations non-conservatrices et rend le degré de non-conservatisme indéterminé. La question de l'effet d'erreurs résultant du rétrocalcul des dimensions des fissures a été examinée, mais il a été constaté qu'il s'agissait d'un effet de second ordre comparé aux « loupés ».

Le CDF des dimensions des fissures détectées ne fournit pas d'informations sur les capacités du système NDI en situation réelle. Le CDF ne donne pas directement les dimensions de fissure détectables avec fiabilité (pour un niveau de confiance donné), mais il donne une première estimation de ces dimensions.

Table of Contents

	Page
Executive Summary	iii
Synthèse	v
List of Figures and Tables	x
Publications of the RTO Applied Vehicle Technology Panel	xi
Task Group Members	xiii
Chapter 1 – Introduction	1-1
1.1 Background	1-1
1.2 Reasons for RTO/AVT Involvement	1-1
1.3 Applied Vehicle Technology Working Group 051	1-2
1.4 Literature Review	1-2
Chapter 2 – Approach for Using NDI in Life Cycle Management	2-1
2.1 Design Criteria (DT versus Safe-Life) Including “Logic Chart”	2-1
2.2 Basis of 90/95 Probability of Detection Requirement	2-3
2.3 Impact on Existing Certification Issues	2-5
2.4 Risk Assessment and POD	2-5
Chapter 3 – Data Collection Process	3-1
3.1 Overview	3-1
3.2 General Summary	3-3
Chapter 4 – Fracture Mechanics	4-1
4.1 Crack Growth Data and Prediction	4-1
4.2 Back-Extrapolation Methodology	4-2
Chapter 5 – Analysis Considerations: POD, CDF and Statistics	5-1
5.1 Introduction	5-1
5.2 POD Model Approach	5-2
5.2.1 Rationale	5-2
5.2.2 Example	5-3
5.2.3 Confidence Limit Calculations for Small Sample Sizes	5-3
5.3 Binomial Model for POD Fitting and Bayesian Safety Level Estimates	5-4
5.3.1 Binomial Model for POD Fitting	5-4
5.3.2 Bayesian Safety Level Estimation	5-5

5.4	Cumulative Distribution Function (CDF) of Detected Cracks	5-5
5.4.1	Rationale	5-5
5.4.2	Example	5-6
5.4.3	Discussion	5-7
5.5	Conclusion	5-8

Chapter 6 – Sensitivity of POD to In-Service Inspection Data **6-1**

6.1	Effect of Undetected Cracks	6-1
6.2	Effect of Crack Size and Sample Size on POD	6-2
6.2.1	POD(a) Model Approach	6-3
6.2.2	CDF of Detected Cracks	6-3
6.3	Effects of Crack Size Errors/Variability on POD Estimation	6-4
6.3.1	Inherent Crack Growth Variability	6-4
6.3.2	Field Measurement Variability	6-5

Chapter 7 – Use of Existing POD Information **7-1**

7.1	Critical Parameters for Data Sets	7-2
7.1.1	Scaling “ a ” vs. “ a ” Response to Pool Data	7-3
7.2	Analysis Checks on Data Compatibility	7-4
7.2.1	Likelihood Ratio Tests	7-4
7.2.2	Wald (Coefficient Standard Error) Tests	7-6
7.2.3	Example Comparison	7-6
7.3	Checks on Assumed POD	7-7
7.4	A Bayesian Method to Pool Data	7-7
7.5	Summary	7-9

Chapter 8 – Conclusions **8-1****Chapter 9 – Recommendations** **9-1****Chapter 10 – Glossary** **10-1****Chapter 11 – References** **11-1****Annex A – Data Collection Process** **A-1****Annex B – Improved Statistical Analysis for Small Data Sets** **B-1****Annex C – Reproducibility and Repeatability in Eddy Current Tests** **C-1****Annex D – Cumulative Distribution Function (CDF) of Detected Cracks** **D-1**

**Annex E – Evaluation of Sample Size, Crack Size and Model in the POD
Characterization of NDI Capability****E-1****Annex F – Distributions of Detected and Undetected Cracks****F-1****Annex G – Crack Size Errors – Nature and Effect on POD Estimation****G-1****Annex H – Binomial and Bayesian Analysis****H-1****Annex I – Simulations of the Effect of Missing Misses in POD Estimation
from In-Service Data****I-1****Annex J – An Approach to Estimating POD Capabilities from Maintenance
Inspection Procedures****J-1**

List of Figures and Tables

Figures		Page
Figure 2-1	A Schematic Illustration of Safe-Life and Damage Tolerance Approaches to Aircraft Design and Operation	2-2
Figure 3-1	A Typical Probability of Detection (POD) Curve	3-2
Figure 4-1	Back-Extrapolation Methodology to Estimate the Sizes of Missed Cracks	4-2
Figure 5-1	Cumulative Log-Normal POD as Fit to In-Service Inspection Data from an F-16 ASIP Control Point	5-3
Figure 5-2	Percentage of Trials with Q1 and Q2 Lower Confidence Limit Curves Non-Conservative at any Point on the Curve, Plotted as a Function of Sample Size	5-4
Figure 5-3	Mean POD Curve for the “Hit/Miss” Data and CDF Curve for the “Hit” Data of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii	5-7
Figure 6-1	Actual Crack Growth Histories Forced to be Coincident at 30 mm	6-5
Figure 7-1	Comparison of 2 Populations of In-Service Inspection Data from F-16 ASIP Control Point	7-7

Tables		Page
Table 6-1	Slow Crack Growth or Slowly Increasing POD	6-2
Table 6-2	Fast Crack Growth or Sharply Increasing POD	6-2

Publications of the RTO

Applied Vehicle Technology Panel

MEETING PROCEEDINGS (MP)

Equipment for Personal Protection (AVT-097) / Personal Protection: Bio-Mechanical Issues and Associated Physio-Pathological Risks (HFM-102)
MP-AVT-097, December 2004

Vehicle Propulsion Integration
MP-AVT-100, August 2004

Novel Vehicle Concepts and Emerging Vehicle Technologies
MP-104, April 2004

Fire Safety and Survivability
MP-103, October 2003

Advances in Rocket Performance Life and Disposal
MP-091, October 2003

Combat Survivability of Air, Sea and Land Vehicles
MP-090, September 2003

Reduction of Military Vehicle Acquisition Time and Cost through Advanced Modelling and Virtual Simulation
MP-089, March 2003

Advanced Flow Management: Symposium Part A – Vortex Flows and High Angle of Attack for Military Vehicles / Part B – Heat Transfer and Cooling in Propulsion and Power Systems
MP-069(I), February 2003

Low Cost Composite Structures / Cost Effective Application of Titanium Alloys in Military Platforms
MP-069(II), February 2003

Ageing Mechanisms and Control: Symposium Part A – Developments in Computational Aero- and Hydro-Acoustics / Part B – Monitoring and Management of Gas Turbine Fleets for Extended Life and Reduced Costs
MP-079(I), February 2003

Ageing Mechanisms and Control: Specialists' Meeting on Life Management Techniques for Ageing Air Vehicles
MP-079(II), February 2003

Unmanned Vehicles (UV) for Aerial, Ground and Naval Military Operations
MP-052, January 2002

Active Control Technology for Enhanced Performance Operational Capabilities of Military Aircraft, Land Vehicles and Sea Vehicles
MP-051, June 2001

Design for Low Cost Operation and Support
MP-37, September 2000

Gas Turbine Operation and Technology for Land, Sea and Air Propulsion and Power Systems (Unclassified)
MP-34, September 2000

Aerodynamic Design and Optimization of Flight Vehicles in a Concurrent Multi-Disciplinary Environment
MP-35, June 2000

Structural Aspects of Flexible Aircraft Control
MP-36, May 2000

New Metallic Materials for the Structure of Aging Aircraft
MP-25, April 2000

Small Rocket Motors and Gas Generators for Land, Sea and Air Launched Weapons Systems
MP-23, April 2000

EDUCATIONAL NOTES (EN)

MEMS Aerospace Applications

EN-AVT-105, February 2004

Internal Aerodynamics in Solid Rocket Propulsion

EN-023, January 2004

Intelligent Systems for Aeronautics

EN-022, June 2003

Active Control of Engine Dynamics

EN-020, November 2002

Supercavitating Flows

EN-010, January 2002

Aging Aircraft Fleets: Structural and Other Subsystem Aspects

EN-015, March 2001

Aging Engines, Avionics, Subsystems and Helicopters

EN-14, October 2000

Measurement Techniques for High Enthalpy and Plasma Flows

EN-8, April 2000

Development and Operation of UAVs for Military and Civil Applications

EN-9, April 2000

Planar Optical Measurements Methods for Gas Turbine Engine Life

EN-6, September 1999

TECHNICAL REPORTS (TR)

MEMS Applications for Land, Sea and Air Vehicles

TR-AVT-078, February 2005

Best Practices for the Mitigation and Control of Foreign Object Damage-Induced High Cycle Fatigue in Gas Turbine Engine Compression System Airfoils

TR-AVT-094, December 2004

Shared Test and Joint Qualification of Advanced Paint Removal Technology

TR-AVT-052, July 2004

All Electric Combat Vehicles (AECV) for Future Applications

TR-AVT-047, July 2004

Future Technological and Operational Challenges Associated with the Single Fuel Concept

TR-066, September 2003

Performance Prediction and Simulation of Gas Turbine Engine Operation

TR-044, April 2002

Evaluation of Methods for Solid Propellant Burning Rate Measurements

TR-043, February 2002

Design Loads for Future Aircraft

TR-045, February 2002

Ice Accretion Simulation Evaluation Test

TR-038, November 2001

NATO East-West Workshop on Magnetic Materials for Power Applications

TR-031, August 2001

Verification and Validation Data for Computational Unsteady Aerodynamics

TR-26, October 2000

Recommended Practices for Monitoring Gas Turbine Engine Life Consumption

TR-28, April 2000

Task Group Members

CHAIRMAN

Mr. David Simpson
Structures and Materials Laboratory
Institute for Aerospace Research, NRC
Building M-13, Montreal Road
Ottawa, Ontario K1A 0R6
CANADA
email: david.simpson@nrc.ca

VICE CHAIRMAN

Dipl. Ing. Georg Guenther
European Aeronautic Defence and
Space Company GmbH
Military Aircraft
MT22
GERMANY
email: Georg.guenther@m.eads.net

MEMBERS

BELGIUM

Capt. R. De Crop
Defence Staff
Dept. Material Resources, V/C
rue d'Evere, 1
B-1140 Brussels
email: decrop@mil.be

Prof. J. Vantomme
Royal Military Academy
Dept. of Civil Engineering
Hobbema, 8
1000 Brussels
email: john.vantomme@rma.ac.be

CANADA

Mr. K. McRae
Air Vehicle Research Detachment
National Research Council
Montreal Road, Building M7
Ottawa, Ontario K1A 0R6
email: ken.mcrae@nrc.ca

GEORGIA

Mr. I. Buchukuri
Ministry of Internal Affairs of Georgia
Moscow Avenue 7/1
380020 Tbilisi
email: tavadze@gas.hepi.edu.ge

GERMANY

Dipl. Ing. E. Grauvogl
EADS, Military Aircraft, ML QM24
Rechlingerstrasse, Munich
email: ernst.grauvogl@m.eads.net

ITALY

Major M. Colavita
Chemistry Department of SCV
Airport "M de Bernardi"
00040 Pomezia
email: mcolavita@tiscalinet.it

THE NETHERLANDS

Mr. J. Heida
NLR, POB 153
8300 AD Emmeloord
email: jheida@nlr.nl

PORTUGAL

Lt. J. Nogueira
Comando Logistico e Administrativo
Direccao de Mecanica Aeronautica/GZE
avenue dea Forca Aerea Portuguesa
email: jnogueira@emfa.pt

UNITED KINGDOM

Dr. D. Bruce
DSTL Farnborough
Hants, GU14 0LX
email: dabruce@dstl.gov.uk

Wg. Cdr. R.B. Eckersely
AS1, B. Block
RAF. Wyton
Huntingdon Cambs, PE28 2EA
email: asi@logistics.org.uk

UNITED KINGDOM (cont'd)

Mr. P.M. Edwards
NDT Squadron, Royal Air Force
St Athan, Barry
Vale of Glamorgan, CF62 4JF
email: raf.ndfs@ukonline.co.uk

Mr. S. Spence
MAA/BAe
Warton W4 27C
email: stuart.spence@bae.co.uk

UNITED STATES

Dr. A.P. Berens
University of Dayton Research Institute
300 College Park, Dayton, OH 45469-0120
email: berens@udri.udayston.edu

Dr. Floyd Spencer
Sandia National Laboratories
P.O. Box 5800, MS 0829
Albuquerque, NM 87185
email: fwspec@sandia.gov

CONTRIBUTORS

BELGIUM

Mr. P. Servais
Comd Rsc NDI
Quartier Roi Albert 1
rue de la Fusée 70
B-1130 Brussels
email: Pierre.servais@mil.be

CANADA

Mr. D.S. Forsyth
Institute for Aerospace Research
National Research Council Canada
M-14, 1200 Montreal Road
Ottawa, Ontario K1A 0R6
email: david.forsyth@nrc.ca

CZECH REPUBLIC

Ing. J. Fidralsky
AERO Vodochody a.s.
Vice President, Engineering
250 70 ODOLENA VODA
email: jiri.fidralsky@aero.cz

ITALY

Capt. F. De Paolis
Italian Air Force
CVS Reparto Chimico
Aeroporto Pratica di Mare
00040 Pratica di Mare (RM)
email: aematcnd@tiscaliinet.it

UNITED KINGDOM

Sergeant N.R. Thorpe
RAF NDT Sqn
Royal Air Force St Athan
Barry, Vale of Glamorgan, CF62 4JF
email: raf.ndts@ukonline.co.uk

UNITED STATES

Dr. John (Jack) Lincoln†
Wright-Patterson Air Force Base

Mr. W. Rummel
D&W Enterprises Ltd
8776 W. Mountainview Lane
Littleton, CO 80125 9406
email: wardduane@earthlink.net

Chapter 1 – INTRODUCTION

1.1 BACKGROUND

Inspection reliability is one of the corner stones of the “safety-by-inspection” approach for continuing airworthiness of aging aircraft and of the damage tolerance philosophy adopted by many of the NATO members as the basis for ensuring continued airworthiness. Inspection reliability data, usually in the form of technique threshold data and Probability of Detection (POD) data are essential for deriving inspection thresholds and inspection intervals. Frequency and method of inspection are primary drivers of maintenance costs and weapon system availability, therefore there is pressure to delay onset and reduce frequency. Safety depends on detection of discontinuities such as fatigue cracks before they reach a critical size, therefore there is pressure to be conservative in defining onset and frequency. These competing aspects can only be properly evaluated with representative inspection reliability data.

Most non-destructive inspection (NDI) reliability data available results from dedicated round-robin inspection programs whereby the same samples are inspected by disparate technicians under laboratory type, or in some cases, simulated in-service conditions. These data have been frequently challenged on the basis that the inspection conditions are not representative in terms of environment, access and human factors of the conditions seen in service. Analysis of in-service NDI findings can improve our understanding of the performance of NDI. This greater confidence in NDI reliability would allow more effective use of NDI for life extension.

Significant numbers of service detections are occurring, but at present there is no organized process whereby these data are collected and collated for NDI reliability studies. There is undoubtedly a large existing amount of inspection data that cannot be accessed directly. The extent of this data and its usefulness to the NDI reliability program must be understood, and for this reason, the Workshop QUANTIFICATION OF AIRFRAME INSPECTION RELIABILITY UNDIR FIELD CONDITIONS was held in Brussels, in May 1998. The processes under which this data could be collected must be defined and implemented in practical and cost effective terms. Equally, the analytical methods used to calculate NDI reliability must be defined and validated.

1.2 REASONS FOR RTO/AVT INVOLVEMENT

All NATO countries are re-evaluating their defence needs and future military system requirements. Extension of originally projected service lives and usage of ageing aircraft are a fundamental part of the life cycle management processes of NATO air forces. Thus, reduction of maintenance costs, sustainment of safety levels and maintenance of current fleet readiness levels are extremely desirable goals.

Optimum use of non-destructive inspections offers the prospect of substantial savings in life cycle costs, particularly when used to enable life extension for airframe structures and components on a “safety-by-inspection” basis. To realize these savings, it is necessary to optimize the inspection strategy without compromising aircraft safety and this requires knowledge of the reliability of the inspection techniques employed. Inspection data and analysis methods are essential to the ability to derive inspection thresholds and inspection intervals – elements of every maintenance program for the constituents within a fleet. In cases where the frequency and method of inspection are primary drivers of maintenance costs, the optimal safe inspection interval can be deduced from the desired safety level and the reliability of the inspection technique used. In many other instances, the determining factor is the major maintenance schedule of the aircraft due to

INTRODUCTION

the cost of downtime and disassembly to enable inspection. In these cases, the required safety level and maintenance interval determine the minimum inspection reliability and the most cost effective method can be chosen to meet this criterion.

1.3 APPLIED VEHICLE TECHNOLOGY WORKING GROUP 051

The intent of this Working Group was to evaluate the potential of reducing life cycle costs while ensuring flight safety through the use of real field inspection-based probability of detection data. Specific objectives were:

- Define the detailed processes for collecting and documenting in-service inspection results with due consideration of: 1) the present form and availability of data that could be used to perform probability of detection studies; 2) the data collection processes that should be put in place to collect relevant data from future inspections; and 3) what are the relevant parameters that must be collected for reliability analysis.
- Define approaches for using the NDI reliability data in the life cycle management process (both deterministic and probabilistic approaches).
- Implement a pilot study on selected NDI techniques, using field inspection data from disparate NATO nations to generate POD data.
- Compare and discuss the POD data generated from field data to that generated from other methods.
- By way of case studies, compare the effect on life cycle management practices (inspection onset, inspection interval, probability of failure, etc.) of using the field generated POD data and the POD data from other methods.
- Through reference to the costs and benefits of generating POD data from field data, provide substantiated recommendations for the implementation of a substantive effort to generate NDI reliability data from field inspection results. The recommendations should include a definition of a process by which this data can be collected, analyzed and documented.

1.4 LITERATURE REVIEW

Simpson (1981) first investigated the concept of using back-calculated crack sizes from in-service crack detections as an approach to obtaining data for a POD(a) capability characterization of an inspection system. At that time, the available POD(a) analysis tools required the hits and misses from the inspections to be grouped in ranges. Simpson used a regression analysis approach to modeling POD as a function of crack size, as per Lewis et al. (1978). However, there were an insufficient number of cracks to permit judgment on the validity of the technique.

In a series of papers between 1993 and 1995, Brewer investigated the possibility of using back-calculated missed crack sizes from lap joint inspections to estimate POD(a), Brewer and Mengert (1993), Brewer (1993a, 1993b, 1993c, 1994a, 1994b and 1995). Brewer concluded that: a) better modeling information was needed to properly back-calculate crack sizes; b) POD from the available service data does not sufficiently account for the non-detections; and c) the chances of detecting individual cracks in lap joints are not statistically independent.

Miller (1995) used a survival function approach to fit a three-parameter Weibull distribution to estimate POD(a) from inspections of commercial aircraft. Brewer, Mengert and Disario (1996) applied Miller's

analysis technique to Japanese maintenance data and concluded that the technique produced overly conservative results and inferior to those obtained from the maximum likelihood estimates of the POD(a) model parameters.

Inspection results from more than 1000 cracks were analyzed by Endoh, et al. (1993) and Asada, et al. (1998). In these analyses, distributions of the detected cracks were stratified by various factors that could influence the inspection process. No attempts were made to back-calculate crack sizes or to characterize the POD capability of the inspection process in terms of POD.

To avoid the need to assume a specific functional form for POD(a) and to obtain estimates of POD from smaller sample sizes, Bruce (1998) used the binomial distribution to model POD in terms of the proportion of finds to finds plus back-calculated misses. Bayesian inference was used to characterize the POD parameter of the binomial model. This approach is further discussed in Section 5.3.

Heida and Grooteman (1998) used two approaches to the characterization of EC inspection efficacy, using in-service inspection results and the back-calculation of missed crack sizes. The parameters of the POD(a) model were estimated using maximum likelihood (see Section 5.2). This characterization of inspection capability was compared to the sample cumulative distribution function of the detected crack sizes. This approach to NDI characterization is discussed in Section 5.4.

Leemans (1998) conducted a set of analytical/Monte Carlo studies to investigate the effects of back-calculating crack sizes over various intervals, uncertainty in crack sizing and uncertainty in airplane operational usage on the estimate of POD(a). All were judged to have potentially significant effects on the estimate of POD. In subsequent studies, Forsyth et al. (1999) and Leemans (2000), the practical application of the method was investigated. The studies concluded that the approach must be implemented as part of the maintenance planning to ensure consistency of application of the NDI procedure, generation and updating of the necessary crack propagation information, and adequate sizing of the detected cracks. Leemans and Forsyth (2004) subsequently developed a method for updating assumed POD estimates with field inspection data using Bayesian methods.

INTRODUCTION



Chapter 2 – APPROACH FOR USING NDI IN LIFE CYCLE MANAGEMENT

2.1 DESIGN CRITERIA (DT VERSUS SAFE-LIFE) INCLUDING “LOGIC CHART”

The governing specification during the development phase of a new military aircraft is generally found in the Weapon System Design Process Specification (WSDPS) or the analogous document of the country in question, listing performance requirements and expected usage data for the overall system, as well as operational and maintenance requirements for individual components of the structure and its equipment, together with references to methods and means of how to comply with these requirements.

The specification should mirror the customer requirements for the weapon system over the envisaged usage timeframe, generally a period of 20 to 30 years of operation. Considering the typical development periods for modern aircraft systems of 10 to 15 years, the importance of this document is imminent, but also the problem of predicting long-term operational scenarios and maintenance concepts three to four decades in advance becomes obvious.

Structural weight of the aircraft is minimized using advanced design principles and high-strength materials, as well as detailed analysis for the primary and secondary structure to ensure effective usage of the material, and at the same time, adequate fatigue life for the expected service application, leading to today's increased usage of lightweight integral components like machined skins and frames made from high-strength aluminum or titanium alloys for wing and fuselage structures.

Integrating more individual elements into a single piece of structure by eliminating fasteners reduces one primary source for local stress concentrations and typical starting points of cracks and corrosion damage, however, on the other hand it reduces the possibility of large-scale repairs through replacement of sub-components in service, thus increasing the need to use major components like wing skins or fuselage bulkheads for the complete service life of the aircraft.

Engineers use two different approaches for the design of metal aircraft structures to resist fatigue and thus ensure flight safety. These two approaches are “safe-life” and damage tolerance (see Figure 2-1). The so-called safe-life approach is a probabilistic-based method. The safe-life of a structure is that usage period in flight hours when there is a low probability that the strength will degrade below its design ultimate value due to fatigue cracking. The determination of the safe-life of an aircraft depends primarily on the results of a full-scale fatigue test of the structure. The number of simulated flight hours of operational service successfully completed in the laboratory is the “test life” of the structure. The safe-life also depends on the expected distribution of failures. The distribution of failures provides the basis for factoring the test life. The factor is called the “scatter factor.” The distribution of failures may be derived from past experience from similar aircraft or from the results of design development testing preceding the full-scale fatigue test. The test life is divided by the scatter factor to determine the safe-life. The scatter factor (usually in the interval from two to four) is supposed to account for material property and fabrication variations in the population of aircraft. A safe-life design theoretically requires no inspections during the design life.

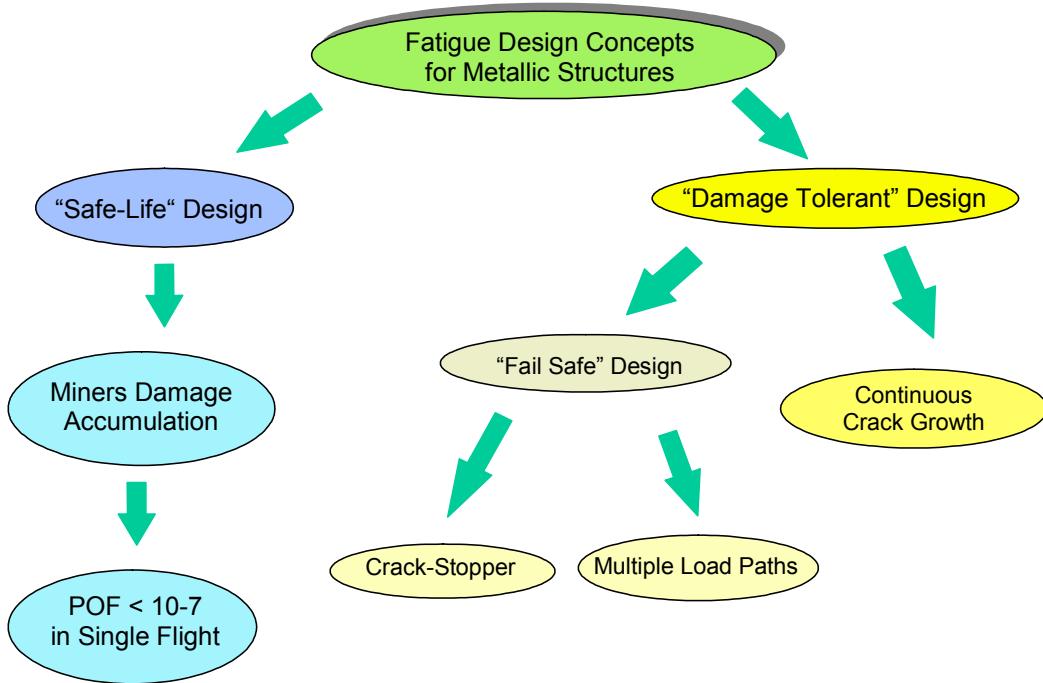


Figure 2-1: A Schematic Illustration of Safe-Life and Damage Tolerance Approaches to Aircraft Design and Operation.

Damage tolerance is the attribute of a structure that permits it to retain its required residual strength for a period of usage without repair, after the structure has sustained specified levels of fatigue, corrosion, accidental or discrete source damage.

The damage tolerance approach as implemented by the USAF is primarily a deterministic method. However, many aspects of it are based on probabilistic methods. The most notable of which, perhaps, is the “inspectable” flaw size. The inspectable flaw size is the size defect that will be detected with a specified probability and confidence. The USAF uses 0.90 probability and 95 percent confidence for their inspectable flaw criterion. A safe-life design typically requires no inspections during the design life. Damage Tolerance designs are inspected periodically at defined spots according to the level of criticality.

One problem with the safe-life approach is that it may not (and did not historically) preclude the use of low ductility materials operating at high stress levels during design loads. To avoid this problem, some authorities use a hybrid approach. That is, they use damage tolerance principles for material selection and safe-life to assure safe operations of their aircraft. Another issue with the safe-life approach is life extension. When the safe-life is reached, there is no easy method of extending the life of the structure without proving the structural integrity with extended full-scale testing, and potential resulting modification of the structure and a full-scale test of the modifications.

Another problem with the safe-life method is that the test life determination is subject to interpretation. Based on the definition of safe-life given above, the allowable flaw size to still retain ultimate load capability may require a fracture analysis of the results of the teardown inspection to determine when this point in the testing occurred.

In addition to the difficulty in the determination of the safe-life, is the difficulty in defining the appropriate scatter factor. Experience has shown that the distribution of failures is dependent on the material used. Ductile aluminum has a narrow failure distribution with a Weibull shape number in the range of four to six. For high-strength steel, however, the Weibull shape number is in the range of two to three, giving it a rather broad failure distribution.

The introduction of damage tolerance principles by some authorities in their structural inspection program reduced the occurrence of fatigue failures in their aircraft. The basis for the process is to assume the structure has a sharp crack at the time it enters service; that is the least upper bound of the expected flaw distribution. The operator makes inspections, with a technique with a quantified reliability, such that the crack is detected before it reaches the point of rapid propagation. Once the crack is detected, the usual procedure is to repair the damage or make a modification and adjust the inspection program accordingly.

The damage tolerance approach is in a state of continual improvement because research and development has lead to better methods in fracture mechanics methods and stress analysis over the last thirty years. There are, however, several concerns about the damage tolerance process. The first is that when a structure that is designed to be fail-safe, the damage tolerance approach may not protect the structure against wide-spread fatigue damage. The reason is that flaws much smaller than those derived from a slow crack growth analysis may cause degradation of the fail-safety of a structure. Consequently, they would not be detected by the NDI procedures used to protect the safety of the intact structure. Another concern is whether all of the critical locations in the structure that need to be inspected have been identified. Most agree that the experience so far has been good. However, when there is inadequate testing or analyses to discover the critical locations, there is a possibility of catastrophic failure of the structure.

2.2 BASIS OF 90/95 PROBABILITY OF DETECTION REQUIREMENT

The 90/95 metric associated with probability of detection (POD) data evolved as a result of the first POD curves produced (see Rummel et al. (1973), Rummel et al. (1974)). The POD curves were produced as a moving average and thus required a large amount of data. The objective of these analyses was to produce data that was consistent with MIL Handbook 5 values for materials properties characterization. For these analyses, there was insufficient data for “A” values and thus “B” values were used. The “B” values constitute a 90% percentile in data output and were used for purposes of plotting the PODs. An estimate of the lower 95% confidence limit can be calculated from sampling tables and was shown on the same POD plots.

In the early 1970's when this work was being performed, deterministic fracture mechanics was the state-of-the art for flaw growth predictions. These required a single-valued “detectable flaw size”. The report was generated for the National Aeronautics and Space Administration (NASA), Lyndon B. Johnson Space Center, and by group agreement (see Castner et al., Rummel (1982)), the point selected for detection capability was that at which the POD curve reached the 90% POD level. This point was termed the “threshold POD limit”. The 90/95 point was that point at which the calculated lower 95% confidence line crossed the 90% POD threshold. In later work using the modeling methods developed by Berens and Hovey (1984), the 90/95 point was simply that point at which the maximum likelihood determined lower 95% confidence bound on the POD curve crosses the 90% POD threshold. If the POD function is considered to be a continuous function, then that method of producing the 90/95 point has been generally accepted.

When the USAF adopted damage tolerance, it was apparent that they must make non-destructive inspections an integral part of the process. This was evident in the drafting of MIL-A-83444, which is the first

APPROACH FOR USING NDI IN LIFE CYCLE MANAGEMENT

specification for damage tolerance used for aircraft design. In this specification, the USAF supposed that an inspection interval of one-quarter of the design life of the aircraft would be acceptable to the logistics community. There was considerable debate, however, about the criterion for establishing the inspectable flaw size. The USAF settled this discussion somewhat arbitrarily. They decided that the inspectable flaw size for establishing the repeat inspection intervals should be that size corresponding to 0.90 probability with 95% confidence. They knew that with a laboratory experiment that they believed to be schedule and cost acceptable, they could establish the inspectable flaw size. Considering that the damage tolerance process was primarily a deterministic method, the USAF had no method of assessing whether this inspectable flaw criterion was adequate.

An opportunity to address this question arose in the early eighties when a service flight loads survey on an USAF trainer showed that there had been a mission change. An earlier damage tolerance assessment was performed for this trainer in Air Training Command usage. This study concluded that the wing center section should be inspected at intervals of 1350 flight hours. This was based on an inspection capability of 2.54 millimeters (corner crack) and an inspection at one half of the safety limit (the time required to grow a crack of 2.54 millimeters to a critical size crack of 5.5 millimeters). In the late seventies, a usage change took place that made the loading environment more severe. A damage tolerance reassessment was made for this new usage, and it was found that, under the same ground rules, the recurring inspection interval should be changed to 430 hours.

This new assessment, which as in the previous assessment used the ninety percent POD for inspections, showed that the inspection frequency should be increased by approximately a factor of three. This, of course, would significantly increase the inspection costs and associated aircraft downtime. To determine if this increased inspection burden was essential to maintain the safety of these aircraft, a risk assessment was performed.

Fortunately, for this trainer, an estimate of crack population could be made from an existing database. Over the preceding several years, destructive teardown inspections have been made on retired trainer wings to provide insight into the possibility of a cracking problem. In all, 19 wings have been torn down and detailed inspections made to quantify the extent of cracking. These examinations revealed that in the critical locations (approximately 100 fasteners or drain holes per wing), roughly 25 percent of the holes had cracks. The upper bound of these cracks was approximately 2.5 millimeters. The USAF believed that these wings were in a state of generalized cracking. Further, they believed that these data were adequate to define the crack population.

For this aircraft, the decision on the adequacy of the criterion for the damage tolerance inspections was made with the knowledge that these wings were in the latter stage of their life. They had operated safely for many years with the inspection program derived from the damage tolerance assessment. The risk assessment results showed that this successful operational experience should have been expected. The risk assessment also showed that inspections are extremely influential in reducing the probability of failure. The problem with this trainer was that a significant population of cracks had grown and were becoming close to critical length in the high-time aircraft. Therefore, the probability had increased that a crack could be missed by the inspection process and become critical. It turned out for this trainer that the 0.9 probability of detection, which was used in the damage tolerance assessment for the new usage, provided an inspection interval of 430 hours. This inspection interval, if used on high-time aircraft, may not adequately protect safety. For this trainer, if the damage tolerance assessment had used a detectable flaw size corresponding to 0.94 probability of detection, then a safe interval would have been provided. In other words, the probability of detection for the damage tolerance assessment required only a relatively small change. It is not known if the results found in this study can be generalized. However, for aircraft that have evidence of significant cracking, the structural engineers

should use all methods available to them to define inspections and/or modifications that will ensure that safety of flight is protected.

In summary, the inspectable flaw size criterion, although arbitrarily chosen, has been successful in maintaining safe operation of aircraft. Care must be exercised in the later stage of an aircraft's life to ensure that generalized cracking is not allowed to increase the risk to an unacceptable level.

It should be emphasized that there is no physical significance to the "90/95" point. Many organizations now commonly use probabilistic fracture mechanics in risk assessments, and these analysis methods in general use the entire POD curve, not a single threshold value. Most often, it is the mean POD curve which is used in these analyses, and confidence curves are not required.

2.3 IMPACT ON EXISTING CERTIFICATION ISSUES

Every aircraft, commercial or military has a maintenance plan that the operator must comply with to maintain its airworthiness certificate. Per definition, a "Safe-Life Design" does not require any inspections during the initial specified design life and usage, consequently any damage showing up during the verification phase needs to be assessed for its relevance in the usage spectrum and requires modification and/or repairs. These repairs and modifications also need to be verified as "safe-life design", either on component basis or, in case major load paths are affected by the modification within the structure through full-scale testing to avoid frequent inspections for modified aircraft.

Therefore, prior to the adoption of the damage tolerance approach, inspections were not a significant part of the maintenance plan. This changed in the seventies when damage tolerance became mandated by many of the certification authorities. For structures designed according to regulations demanding damage tolerance capabilities (i.e. MIL-A83444 and FAR 25), inspections and POD-values for the NDI-method selected are dealt with from the design concept phase onwards, while other regulations allowing safe-life designs with optional damage tolerance elements (i.e. DEF-STAN-00-970 or MIL-A-8866C), do not require inspection reliability assessments for their designs. Since damage tolerant structural analysis does incorporate an initial flaw size in critical structural elements, the capability of the NDI methods to detect these flaws now becomes a certification issue itself, to be demonstrated by test on build-up structures during the fatigue qualification process.

The major airframe manufacturers focused considerable attention on the capability of non-destructive inspections to find cracks from fatigue loading. The findings from this effort then became part of the maintenance plan. For the military, the technical orders were rewritten to identify the specific non-destructive equipment that must be used to perform the inspection. The commercial operators were typically given more freedom of choice. For example, Boeing provided options that could be used to establish inspection intervals for their commercial customers. The intervals depended on the reliability of the specific inspection technique the operator planned to use.

2.4 RISK ASSESSMENT AND POD

Ensuring structural integrity through damage tolerance is most commonly based on deterministic analyses. The growth of the largest, single crack that might be in the most critical location of a structural element is predicted using a sequence of stresses from expected operational use of the aircraft. Maintenance actions for the element are conservatively scheduled from the predicted time for the potential crack to grow to a critical

size. This design philosophy has worked well. However, cracking scenarios can arise in an aging fleet that are not amenable to analyses based on the growth of a monolithic crack. For example, wide-spread fatigue damage can produce complex cracking scenarios in which the structural conditions of the elements in a load path are unknown and conservative assumptions would lead to unacceptable inspection intervals. In these scenarios, structural risk analyses are being used to assess the structural integrity of the load path.

In structural risk analysis, the integrity of a structure is characterized in terms of the single flight probability of failure of the load path. This probabilistic evaluation of strength versus stress is dynamic, since strength degrades as fatigue cracks in the load path grow, and the condition of the structure might change during maintenance actions. In a risk analysis, the condition of the structure is modeled in terms of distributions of the strength limiting cracks at the critical locations, and fracture mechanics tools are used to predict the growth of the distributions of cracks as a function of flight hours. Probability of failure is calculated from the distributions of strength and expected stresses that will be experienced during a flight at time T. Maintenance actions would be scheduled at intervals that provide an acceptably small failure probability. For example, Lincoln (2000) has suggested that 10^{-7} is an acceptable upper bound on single flight failure probability for United States Air Force applications.

While a single crack size with a high detection probability may provide a sufficient description of NDI capability for deterministic crack growth analyses, an estimate of the entire $POD(a)$ function is needed for probabilistic risk assessment. To account for the effect of an inspection with attendant repair when a crack is found, the analytical model must account for both the distribution of the crack sizes in an element at the inspection and the probabilistic nature of the inspection process. Analytically, let $f_{before}(a)$ and $f_{after}(a)$ represent the probability densities of crack sizes in the population of structural elements before and after an inspection with detected cracks being repaired. Let $POD(a)$ represent the probability of detecting a crack of size a and $f_R(a)$ represent the probability density of equivalent flaw sizes at the repaired crack sites. Then

$$f_{after}(a) = Pf_R(a) + [1 - POD(a)]f_{before}(a) \quad (2-1)$$

where P is the expected proportion of all cracks that will be detected and repaired.

$$P = \int_0^{\infty} POD(a)f_{before}(a)da \quad (2-2)$$

If a single crack size, say a_{NDI} , is used to represent an inspection, the analysis assumes that cracks smaller than a_{NDI} are missed during an inspection, while cracks greater than a_{NDI} are found and repaired. For this formulation, $POD(a)$ would be misrepresented as a step function at a_{NDI} . Since such a misrepresentation of the $POD(a)$ function will significantly influence the calculation of failure probabilities, a realistic estimate of the $POD(a)$ function is required for the estimation of structural failure risks.

Examples of the use of risk analysis in airframe structures can be found in Lincoln (1997), Cochran et al. (1991) and Berens et al. (1998). Examples of the use of probabilistic analyses in engine structures can be found in Yang and Chen (1985), Koul et al. (1985), Harris (1987) and Roth (1992).

Chapter 3 – DATA COLLECTION PROCESS

3.1 OVERVIEW

Data requirements for use in developing Probability of Detection (POD) outputs are:

- known crack/artifact sizes,
- rigid calibration control, and
- rigid procedure control.

The usefulness of maintenance data collected is dependent in large part on the fidelity and precision of that data. Non-destructive inspection (NDI) utilizes indirect measurement of a material characteristic or parameter and correlation of that measurement to a desired material characteristic or property. Reliable detection of cracks (or other discontinuities) by an applied (NDI) procedure is dependent on:

- capability,
- reproducibility, and
- repeatability.

The CAPABILITY of a procedure is roughly characterized by the inherent signal and noise responses as applied to a specific test object and crack-to-crack variances within the test object. The capability and hence applicability of an NDI procedure is dependent on the fidelity and precision of the causal model relationship between the measured parameters (NDI output) and the desired characteristic. This is inherent in the physics of the NDI method and application parameters including the threshold limit used for purposes of accept or reject.

The REPRODUCIBILITY of a procedure is generally characterized by the inherent capability and variances in the procedure “calibration” process. Reproducibility is defined as the ability for a specific NDI technique to be performed or “reproduced” from a set of specifications. For example, can one maintenance base reproduce a result (signal output and decision) that is the same as that produced at another base.

The REPEATABILITY of a procedure is generally characterized by process control and variances in application of the procedure and includes “human factors” for those applications involving signal or pattern recognition by human operators. Repeatability is defined as the ability for a specific NDI technique to be used repeatedly on the same specimen and to obtain the same result.

Finally, accuracy and precision in data recording are required to provide confidence in the data provided.

Probability of Detection (POD) methodology was initially developed to assess and validate inherent capabilities of various non-destructive inspection (NDI) procedures. Reproducibility and repeatability were assumed and output variances were attributed to “human factors”. Precision in crack size measurement and documentation was required to minimize variances in NDI output (capability) as a function for crack size. Rigor and confidence in the detection process required a significant number of detection opportunities (trials) to characterize and quantify the detection output. Detection was and is generally recorded as a “HIT OR MISS” (detect or failure to detect) output. The basis for detection (detection threshold) was assumed to be constant. Good engineering practice and economics required that the detection threshold must result in a low level of “false calls” (a detection call when no crack is present).

DATA COLLECTION PROCESS

Probability of Detection (POD) methodology requires passing a large number of cracks or other anomalies (typically 60 or more) through an NDI process and recording the results as “HIT OR MISS”, or as a scalar quantity with respect to actual crack size. The resulting data is then analyzed and fit to a cumulative log-normal model, as is discussed in Section 5.2 of this document. Figure 3-1 shows a typical POD curve.

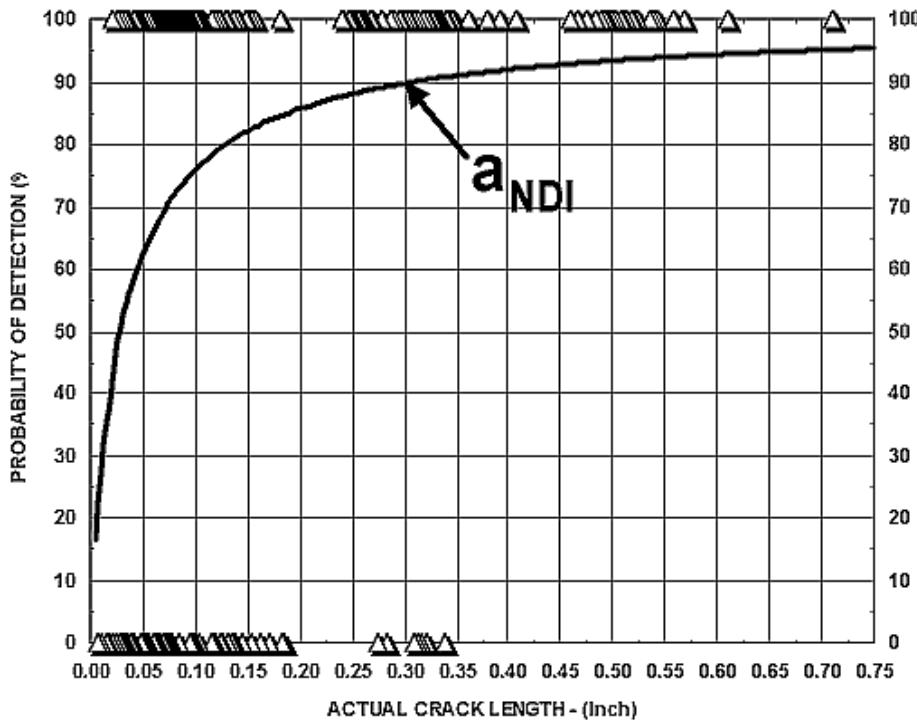


Figure 3-1: A Typical Probability of Detection (POD) Curve.

Wide-spread use of the **POD** methodology to characterize, quantify and validate **NDI** procedure capabilities has identified significant variance in both **REPRODUCIBILITY** and **REPEATABILITY** due to variances in “calibration” and equipment/probe/transducer/inspection materials performance. It follows that the greater the variance in the **REPRODUCIBILITY** and **REPEATABILITY**, the greater the variance in applied NDI procedures and the resultant POD output. This has been one of the key obstacles to acceptance of the POD methodology – experiments for POD estimation must account for the expected variances at the level of implementation of the technique, not just at the laboratory level. Annex C provides an example of variances in reproducibility and repeatability in a practical maintenance situation.

In addition to challenges of variances in **REPRODUCIBILITY** and **REPEATABILITY** in applied **NDI** procedures, **POD** characterization from maintenance data involves additional challenges in precision, in sizing the detected anomalies at the time of the **NDI** procedure application and an absence of crack sizes for “missed anomalies”. The fidelity and usefulness of **POD** performance characterization from maintenance data is therefore dependent on variances in data quality (variance bounds). Variance in the quality of recorded data may result in variances in POD that neither reflect an accurate or useful capability of an **NDI** procedure.

An in-depth review of the technical challenges associated with the data collection process is included in Annex A.

3.2 GENERAL SUMMARY

NDI data from maintenance records is known to vary considerably in quality and fidelity. Variances in some maintenance data is known to be such that POD analysis is not possible or useful. Judgment must be exercised in selecting those data sets that can provide useful POD analysis output. Rigorous and/or judgmental analysis of variances associated with data collection and documentation may be used for initial screening.

If POD analysis from maintenance data is desired or required for future applications to fleet structure integrity analyses, fleet maintenance and aircraft life extensions, additional rigor in NDI procedure validation, calibration, application and documentation are required. The discussions provided in Annex A provide overall guidelines for data quality requirements. Structural integrity analysis needs are expected to drive requirements for improved NDI data quality and documentation. The result will be not only more useful data, but improved inspection reliability by reduction of variances in the NDI data acquisition and reporting process.

DATA COLLECTION PROCESS



Chapter 4 – FRACTURE MECHANICS

4.1 CRACK GROWTH DATA AND PREDICTION

According to W. Schütz (1996), the initial concepts of fracture mechanics started with the Englishman Griffith in 1922. In 1958, G. Irwin from the US Navy enlarged on the ideas of Griffith. He recognized that the stress intensity factor $K = \sigma \sqrt{pa}$ (σ is the stress and a is the crack length) was a means for determining the static strength of a material in the cracked condition. If K reaches the fracture toughness of the material, a spontaneous failure occurs. This was the beginning of linear elastic fracture mechanics. In 1962, P. Paris in his dissertation postulated that the crack extension in a single cycle of loading was proportional to the n^{th} power of the change in stress intensity. The number n is typically in the range of three to five. The work of Paris opened the door to crack growth calculations for structures in the environment repeated loading.

To make the crack growth calculations, the researchers needed the crack growth function. This is the crack growth per loading cycle as dependent on the change in stress intensity for the loading cycle. These functions were experimentally derived in laboratories under various environmental conditions representative of the aircraft environment. They are usually developed using “compact tension” specimens or “center crack” specimens. The effects of R (the ratio of the maximum to minimum stress in constant amplitude loading), geometry, temperature, grain direction and grain size influence the crack growth rates and therefore must be included in a test program.

There is typically a significant amount of scatter in the crack growth functions because of experimental errors and the variability in the material properties. The USAF approach is to use the average of no less than three specimens for the fracture mechanics calculations. However, the effect on crack growth rate from material variability is significant enough to influence the assessment of NDI reliability from operational experience.

The experimental procedures used to develop the crack growth functions have led to a problem in some materials because of the “crack closure” effects. These effects caused a premature reduction in crack growth rate, as the ΔK became smaller. This is called the “long crack threshold.” It is important for applications to engines and rotating components in helicopters, since the stress intensities in these structures are typically smaller than those found for fixed-wing aircraft components. This long crack threshold anomaly may have profound influence on inspection intervals.

For the damage tolerance assessment, it is the intent in the derivation of the stress spectra to determine the “baseline usage” as an average usage for the force. Consequently, most of the information derived during the development of an aircraft is for the baseline spectra. Sensitivity studies are also conducted to ensure that the tail number tracking can be accomplished with acceptable accuracy.

The spectrum for a commercial aircraft, particularly the large category transport aircraft, changes little during its life. This is not the case for military aircraft. Even the large cargo carrying aircraft undergo significant changes in their spectra of loading. One reason for this is change in tactics. To avoid detection by radar, these aircraft at times will fly at low altitudes. When this happens, the crack growth rates may be influenced by an order of magnitude. Change in tactics may also cause large seldom-occurring loads with attendant retardation effects. Retardation can easily affect the crack growth rate by a factor of more than two.

Training can also have a major effect on the usage spectrum. Since the training environment requires numerous touch-and-go landings, the distribution of damage to the airframe components will be different than that found in normal operations.

As an aircraft ages, it usually has an increase in its mass because of new equipment being added. Experience with high-performance aircraft in the USAF indicates that this effect can affect the crack growth rate by a factor of three or more.

Tracking of each aircraft in the inventory enables the operator to compensate for usage changes by modification of the inspection intervals.

4.2 BACK-EXTRAPOLATION METHODOLOGY

NDI systems are generally classified into two categories depending on the outcome of an inspection: NDI systems producing only qualitative information such as the presence or absence of a crack indication ("hit/miss" data), and NDI systems recording a signal response (\hat{a}) that correlates with the actual size (a) of the indicated crack (" \hat{a} vs. a" data). Probability of detection (POD) curves can be calculated for both NDI systems. Most in-service inspections of aircraft, however, do not record the signal response. Therefore only the analysis of "hit/miss" data will be considered in this report.

Field inspection data generally comprise hit data only, i.e. a registration of cracks found during scheduled inspections. Information of the sizes of undetected cracks (misses), however, is necessary for the construction of a POD curve. Cracks that were detected during a scheduled inspection were possibly missed at previous inspection times. An estimation of the sizes of these missed cracks can be done when a crack growth curve is available for that inspection configuration. This back-extrapolation methodology is illustrated in Figure 4-1.

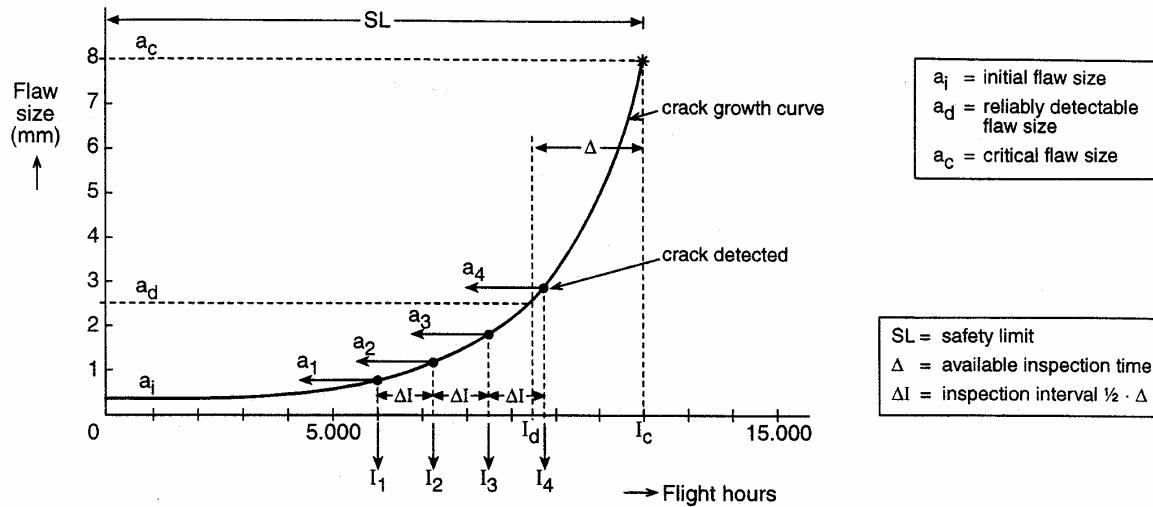


Figure 4-1: Back-Extrapolation Methodology to Estimate the Sizes of Missed Cracks.

Figure 4-1 shows the crack growth curve assumed valid for a particular inspection configuration. In accordance with damage tolerance design philosophy, the initial inspection time I_1 is scheduled at $SL/2$, i.e. half the crack growth time from initial crack size (a_i) to critical crack size (a_c). The inspection interval

ΔI is $\frac{1}{2} \Delta$, i.e. half the crack growth time from reliably detectable crack size (a_d) to critical crack size (a_c). Assume that a crack has been detected for the first time at the fourth scheduled inspection (I_4); the size of the crack at that time was a_4 . This implies that smaller cracks have been missed at the previous three scheduled inspection times I_3 , I_2 and I_1 . The sizes of these missed cracks can be estimated using the crack growth curve resulting in the values a_3 , a_2 and a_1 , respectively. The crack detection in this example hence provides one hit of size a_4 and three misses of sizes a_3 , a_2 and a_1 as input data for the “hit/miss” database of this particular inspection configuration. In practice, the size of the detected crack will not be exactly a_4 , but can have a different value. Back-extrapolation of the sizes of missed cracks at the previous scheduled inspection times will then of course start at that specific crack size value on the crack growth curve.

The reliability of the back-extrapolation methodology depends on the accuracy of the sizing of the detected crack and on the validity of the crack growth curve. Prominent factors in the crack growth curve are the uncertainty in the material parameters of the crack growth equation and differences in spectrum severity for different aircraft. These aspects will be discussed further in Chapter 6.



Chapter 5 – ANALYSIS CONSIDERATIONS: POD, CDF AND STATISTICS

5.1 INTRODUCTION

While NDI systems are capable of finding “small” cracks, ensuring safety through damage tolerance is based on the largest crack that might be in the structure after an inspection. Thus, the focus of NDI capability evaluation for damage tolerance is the largest crack that might be missed at an inspection. NDI techniques do not always produce a correct indication when applied by inspectors to cracks of the same size. The ability and attitude of the operator, the geometry and material of the structure, the environment in which the inspection takes place, and the location, orientation, geometry and size of the crack all influence the chances of detection. When considering the detection efficacy of an NDI system as a function of only crack size, ignoring other factors adds to the uncertainty of crack detection at the small sizes of interest. This uncertainty is quantified in terms of the probability of detection (POD) of cracks of a fixed size, a . POD(a) is defined as the proportion of all cracks of size a that will be detected by the NDI system when applied by representative inspectors to the population of structural elements in a defined environment.

Estimating the POD capability of an NDI system for a specific application requires statistical analysis of the results of inspections for which the sizes of the cracks that were both detected and missed are known. Traditionally, such inspection results have been obtained from controlled experiments using specimens with cracks of known size and location. Because of the artificial nature of both the specimens and the inspection conditions, POD capabilities estimated from such experiments were generally considered to be optimistic to an unknown degree.

This study addresses the use of in-service inspection results in which multiple inspections of a structural element have been performed using the same inspection procedure. Cracks that have been detected were missed at the previous inspection times. Given valid crack projection data, the sizes of the misses at the previous inspections can be calculated. This back-growth calculation was presented in Chapter 4. The measured sizes of the detected cracks and the calculated sizes of the missed cracks are data from which a POD capability characterization can be calculated. However, it must be noted that such estimates of POD are most likely to be biased in a non-conservative direction. The crack sizes for missed cracks can be estimated only from the cracks that were detected. In a realistic inspection scenario, the relative frequency of crack sizes in the population will be decreasing over the crack sizes for which the POD(a) is increasing. There will be a range of crack sizes for which misses are more likely than finds, and these misses will not be represented in the analysis. Reasonable inspection scenarios can be envisioned for which the number of missed misses will exceed the number of finds. The exclusion of these missed misses will lead to a non-conservative estimate of POD. This topic is extensively addressed in Section 6.1.

This section of the report presents three analysis approaches to characterizing the inspection capability from in-service crack detections. The distinct approaches are:

- 1) using a model for the POD(a) function with maximum likelihood estimation of the parameters of the function and confidence bound on POD(a),
- 2) using a binomial model as the basis to characterize POD, and
- 3) using the cumulative distribution function (CDF) of detected cracks as an indication of inspection capability.

5.2 POD MODEL APPROACH

The POD(a) model approach to characterizing the capability of an inspection system comprises the assumption of a functional form for the POD(a) model, the estimation of the parameters of the model, and quantifying the uncertainty in POD estimates by confidence limits. Extensive development and discussion for this approach to POD characterization can be found in Berens (1988), Petrin et al. (1993) and Forsyth and Fahr (1998). The following presents a brief summary of the method for the type of in-service inspection data that would be available for analysis and the results of an example application.

5.2.1 Rationale

A reasonable assumption is that the chances of crack detection will increase over the range of crack sizes of interest. There are many equations that can model such increasing probabilities and no single equation is best for characterizing all inspection reliability scenarios. Because of the scatter in chances of detecting different cracks of the same size, it is reasonable to select a common model for analyzing POD as a function of only crack size. In Berens and Hovey (1981), data from multiple inspections of airframe components were used to compare seven different equations for POD(a). This study concluded that the cumulative log-normal and log odds equations provided as good or better models than the other six that were considered. This conclusion has been supported by analysis of inspection response data from eddy current inspections, wherein the observed distribution of responses about a mean response led to a cumulative log-normal model for the POD function, Berens (1988). Because no other single equation has been shown to be more universal, the cumulative log-normal model has evolved into the most commonly used model for aircraft applications.

The cumulative log-normal equation for the POD(a) functions is:

$$\text{POD}(a) = \Phi[(\ln a - \mu)/\sigma] \quad (5.1)$$

where $\Phi(z)$ is the standard normal cumulative distribution function. The parameter μ is the natural logarithm of the crack size for which there is 50% detectability. The parameter σ is a scale parameter that determines the flatness of the POD function – smaller σ yields steeper POD functions. The parameters μ and σ are estimated from the inspection results of cracks of known size.

Damage tolerance analyses are driven by the single crack size characterization of inspection capability for which there is a high probability of detection. Typically, the one number characterization of the capability of the NDI system is expressed in terms of the crack length for which there is 90% probability of detection. Denote this crack size by a_{90} . For the cumulative log-normal POD function,

$$a_{90} = \exp(\mu + 1.282 \sigma) \quad (5.2)$$

But a_{90} can only be estimated and there is sampling uncertainty in the estimate. To cover this variability, an upper confidence bound can be placed on the best estimate of a_{90} . The use of an upper 95% confidence bound, the $a_{90/95}$ crack size, has become the de facto standard for this characterization of NDI capability. Safety factors are usually applied to inspection intervals in order to preserve conservatism in risk.

Inspection results are recorded in two distinct formats. The format that will be most commonly available from the in-service inspection database expresses the results only in terms of crack size and whether or not the crack was detected. Such data are known as find/no find, “hit/miss”, or pass/fail data. The dichotomous inspection results are represented by the data pair (a_i, Z_i) , where a_i is the size of crack i and Z_i represents the

outcome of the inspection of crack i : $Z_i = 1$ for the crack being found (hit or pass) and $Z_i = 0$ for the crack not being found (miss or fail). Maximum likelihood estimates of the parameters of the $POD(a)$ model are obtained from the (a_i, Z_i) data. Asymptotic properties of the maximum likelihood estimates are used to calculate the confidence bound on the estimate of the crack sizes of interest, say a_{90} (see Petrin et al. (1993) or Berens (1988)).

5.2.2 Example

As an example of the model approach to $POD(a)$ analysis, the cumulative log-normal model was fit to the in-service inspection data of a control point on the F-16 airframe. The data used in the analysis comprise 39 detected cracks with the sizes of 51 misses as calculated using individual aircraft crack growth severity spectra. The data are listed in Table D-3 of Annex D. The resulting $POD(a)$ function and the 95% confidence bound for individual $POD(a)$ values are shown in Figure 5-1. The cracks that were detected at the in-service inspections are plotted at $(a, 1)$ where a is the measured size of the detected crack. The misses are plotted at $(a, 0)$ where a is the estimated crack size at previous inspections.

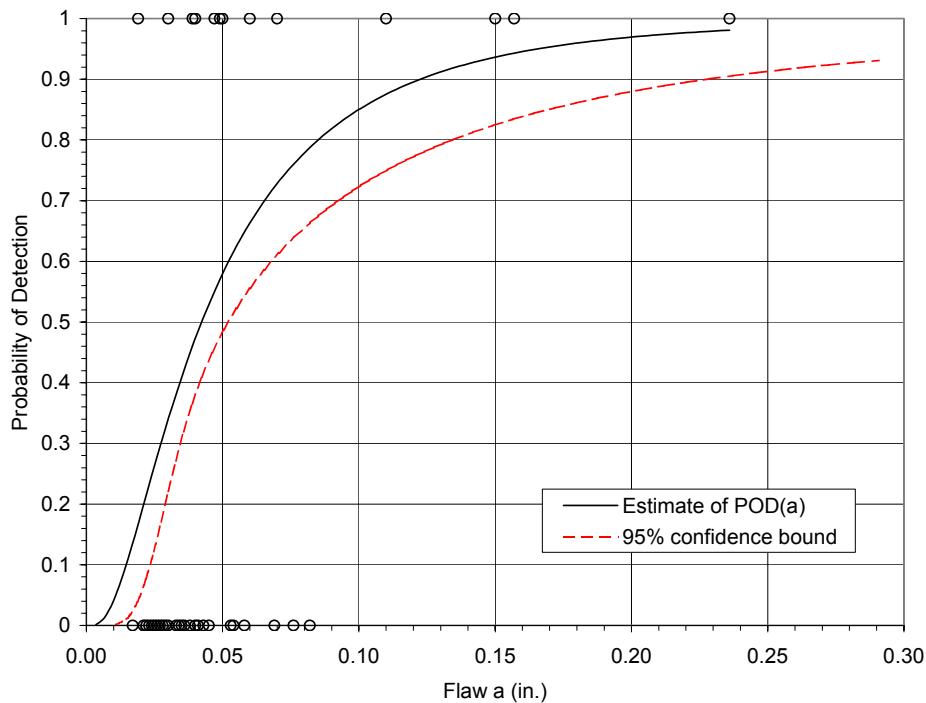


Figure 5-1: Cumulative Log-Normal POD as Fit to In-Service Inspection Data from an F-16 ASIP Control Point.

5.2.3 Confidence Limit Calculations for Small Sample Sizes

In the situations where the “90/95” discontinuity size is used to calculate inspection intervals and/or risk, it has been shown by Harding and Hugo (2003) that methods previously reported in the literature for the calculation of 95% confidence levels can become overly conservative for small sample sizes. An example is shown in Figure 5-2, where the data denoted Q1 is calculated using the method described in USAF MIL-HDBK-1823, and the data denoted Q2 is calculated by a method described in Harding and Hugo (2003).

Below sample sizes of about 100, the calculation of MIL-HDBK-1823 becomes relatively non-conservative. Further details on these calculations are provided in Annex B, courtesy of the original authors.

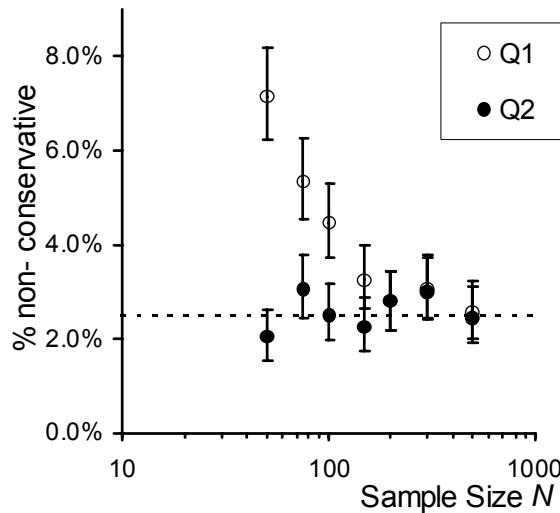


Figure 5-2: Percentage of Trials with Q1 and Q2 Lower Confidence Limit Curves Non-Conservative at any Point on the Curve, Plotted as a Function of Sample Size (see Harding and Hugo (2003)). Figure reproduced with the permission of the authors.

5.3 BINOMIAL MODEL FOR POD FITTING AND BAYESIAN SAFETY LEVEL ESTIMATES

5.3.1 Binomial Model for POD Fitting

The original work in POD was performed using a framework of binomial data analysis. In this model, hits and misses are grouped or “binned” into ranges, where each bin is then given a mean POD assuming a binomial distribution within the range. A brief description of an implementation of the range interval method for fitting POD curves to inspection data is given in the following text.

In the range interval method, it is assumed that the variability of POD within a small crack size range or interval is small and the detection within that range follows a binomial distribution (see Berens and Hovey (1982)).

To implement the range interval method, the crack data is divided into t intervals of equal length. The probability of detection is calculated for each interval as being the ratio of cracks detected to the total number of cracks in that interval. This gives t data points. The t data pairs of POD and crack length are transformed into a linear domain, and a linear regression is performed on the data pairs in order to obtain the intercept and slope parameters, α and β , of the log-logistic function (equation 5.3). The reverse transformation gives the POD curve. The functional form of the log-logistic distribution is as follows:

$$P_i = \frac{\exp(\alpha + \beta \ln(a_i))}{1 + \exp(\alpha + \beta \ln(a_i))} \quad (5.3)$$

where P_i is the probability of detection for crack i , a_i is the length of crack i , and α and β are constant parameters which define the curve. The data points are transformed into a domain where the POD relationship is linear, using the following transformations on the log-logistic distribution function:

$$Y_i = \ln\left(\frac{p_i}{1 - p_i}\right), \quad X_i = \ln(a_i) \quad (5.4)$$

where p_i is the proportion of cracks detected and a_i is the crack length in the interval i . The regression analysis is applied only to the data intervals up to and including the first of any three consecutive intervals where the proportion of cracks detected is 100%. The result of the transformation on equation 5.3 is a set of points which are fitted with the line:

$$Y = \alpha + \beta X \quad (5.5)$$

These parameters α and β can be substituted into equation 5.3 and used to calculate a POD curve for a range of crack lengths.

A number of methods have been used to place confidence bounds on POD curves estimated using RIM. The binomial assumption results in confidence bounds that are highly dependent on the number of cracks in the interval of concern, and 95% confidence curves are often very conservative in comparison to those calculated using the method of Section 5.2 (see Fahr et al. (1993)).

5.3.2 Bayesian Safety Level Estimation

In the Bayesian approach, the degree of confidence in a particular outcome before an experiment is expressed as a “prior” distribution of probabilities. In applying the approach to NDI reliability assessment, the prior distribution is chosen as the level of confidence in achieving given values for the probability of detection. An initial experiment is then carried out. The outcome of the initial experiment is used to update the prior distribution, producing a “posterior” distribution reflecting the revised degree of confidence in the possible outcomes as a result of including the additional information which has been obtained. Bayesian confidence levels and intervals can be estimated from the posterior distribution which can be used to determine the effectiveness of the inspection. Finally, the Bayesian analysis can be used to produce a third distribution, the “predictive” distribution, which is calculated directly from the posterior distribution. This gives the probability of any outcome in a subsequent experiment, given the initial level of knowledge in the prior distribution and the additional information from the initial experiment.

In Annex H, a Bayesian method is developed and demonstrated to estimate safety levels for situations where safety is maintained by inspections, based on a binomial model for inspection reliability. The performance of the Bayesian method in estimating safety is found to be less conservative than that of the range interval method.

5.4 CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF DETECTED CRACKS

5.4.1 Rationale

In-service inspection of aircraft generally yields information about the detected crack size only. When crack growth data are available for each crack detected, the missed crack sizes during previous inspections can be

estimated using the back-extrapolation methodology discussed in Chapter 4.2. The result is a database of the “hit/miss” type, and a POD curve can be constructed. Two different statistical methods can be used for this purpose, viz. binomial or curve fitting methods. In Fahr et al. (1993), it was concluded that the curve fitting method with the log-normal distribution function provides the most realistic POD results.

When crack growth data are **not** available, as is often the case for in-service aircraft inspections, it is not possible to estimate the “miss” data anymore, and a POD curve cannot be constructed from the available “hit” data only. However, the crack detection data can still be used to yield information about the in-service detectable crack size by means of a detection threshold histogram, Simpson (1981). For this purpose, the available data are grouped in appropriate intervals of detected crack size and a histogram is made of the frequency of detection versus crack size. The histogram can give information such as the sensitivity of inspection (detection threshold) and the mean crack size detected.

A further approach is to assume a Probability Density Function (PDF) for the crack sizes detected and to calculate its integral, i.e. the Cumulative Distribution Function (CDF), Heida and Grooteman (1998). By analogy with standard POD calculations with both “hit” and “miss” data available, a log-normal function can be assumed for the crack sizes detected (“hit” data a_i):

$$\text{PDF} : f(a) = \frac{1}{a \sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\ln(a) - \mu}{\sigma} \right)^2} \quad (5.3)$$

where the parameters μ and σ are the mean (location parameter) and standard deviation (scale parameter) of the log crack sizes detected. These parameters can be determined with a parameter estimation procedure such as the Maximum Likelihood Estimators (MLE) method or the least squares method.

Next, the CDF can be calculated by taking the integral of the PDF, indicating the probability that the detected crack size has a value less than or equal to a_i :

$$\text{CDF} : F(a_i) = \int_{x=0}^{x=a_i} f(x) dx \quad (5.4)$$

The CDF-hits curve can provide information about the detectability of cracks in a field inspection environment. The differences between POD(a) and the CDF of detected cracks are illustrated in Annex F.

5.4.2 Example

To illustrate the PDF/CDF approach, the inspection data of an AGARD round-robin NDI demonstration program and the in-service inspection data of a control point of the F-16 airframe structure have been reviewed (see Fahr et al. (1995) and Heida and Grooteman (1998), respectively). Annex D gives the analysis of these data, together with the corresponding POD and CDF curves.

An example of the PDF/CDF approach is given in Annex D for the inspection data of the F-16 centre fuselage longeron (see Heida and Grooteman (1998)).

The longeron is a tee-extrusion machined from 2024-T62 aluminium whose purpose is to distribute flight loads from the fuselage upper skin to the centre fuselage structure. High positive g-loads cause fatigue cracking in the tab radii of the longeron. NDI of the tab radii involves a manual eddy current inspection

technique using standard phase analysis equipment and standard eddy current probes. The database of Heida and Grooteman (1998), status March 1998, comprises 28 “hit” data points and 36 “miss” data points back-extrapolated using a durability crack growth curve. The 28 “hit” data points have been used to calculate a CDF-hits curve and the 64 “hit/miss” data points have been used to calculate a mean POD curve (Figure 5-3). For the curves, a log-normal distribution function was assumed. The location (μ) and scale (σ) parameters were determined with the least-squares method (CDF curve) or with the MLE method (POD curve), resulting in (μ , σ) values of (1.4, 1.1) mm and (1.2, 1.0) mm, respectively.

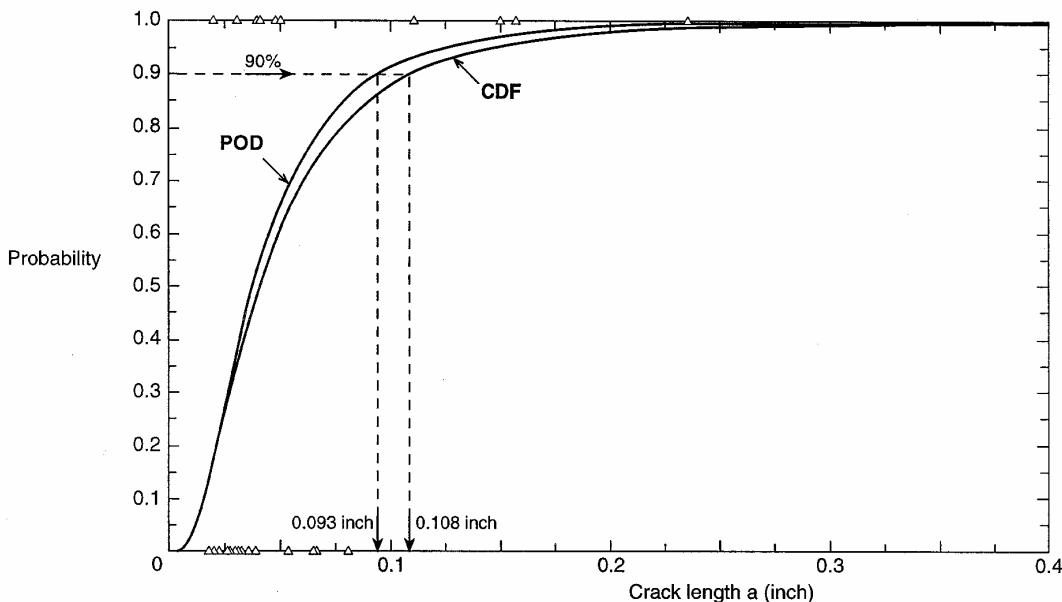


Figure 5-3: Mean POD Curve for the “Hit/Miss” Data and CDF Curve for the “Hit” Data of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii (see Heida and Grooteman (1998)).

The POD and CDF curves correlate remarkably well, with the CDF-hits curve located slightly to the right of the mean POD curve, i.e. it is conservative. An arbitrary 90% probability criterion yields the crack lengths of 2.7 mm (0.108 inch) and 2.4 mm (0.093 inch) for the CDF-hits and POD curve, respectively. It is emphasised that these values cannot be compared directly; 2.4 mm is the crack length for which there is a 90% probability of detection (confidence level 50%), while 2.7 mm is the crack length for which there is a 90% probability that the detected cracks have a length less than or equal to 2.7 mm. For this inspection case, the CDF-hits curve gives a conservative estimate of the detectable crack length, here arbitrarily defined as the crack length for which there is a mean POD of 90%. For other inspection cases described in Annex D, however, the CDF-hits curve is not necessarily located conservatively with respect to the POD curve.

5.4.3 Discussion

Information about the detectability of cracks in a field inspection environment can best be obtained with POD curves constructed from “hit/miss” data sets. However, the analysis in Annex D shows that it will be very difficult in practice to produce a “reliable” POD curve. This is caused by unreliability in the values of the detected crack sizes, by unreliability in the values of the “miss” data (back-extrapolation procedure in general) and because even small changes in the “hit/miss” data set can have a large influence on the POD curve.

Further, for many inspection cases it will not be possible even to construct a “hit/miss” data set, for example in the absence of crack growth data, so that “miss” data points cannot be determined. In those cases, the CDF-hits curve can be of use. This curve is quite stable and less vulnerable to changes in the data set than the POD curve. It is emphasised that the CDF-hits curve is not a POD curve, but it does provide information about the detectability of cracks in a field inspection environment. Furthermore, it can give a first estimate of the detectable crack size.

5.5 CONCLUSION

The CDF-hits curve has a shape similar to the POD curve. It is **not** the POD curve, but it does provide information about the detectability of cracks in a field inspection environment. The CDF-hits curve does not directly yield the reliably detectable crack size (at a given confidence level), but it gives a first estimate of this size.

Chapter 6 – SENSITIVITY OF POD TO IN-SERVICE INSPECTION DATA

There are several sources of potential bias and variability in the in-service inspection data that could influence the estimate of POD. The results of studies into the effects of these potential sources of errors are summarized in this chapter. Details of the studies are presented in the Annexes.

6.1 EFFECT OF UNDETECTED CRACKS

Probability of detection for a single crack condition would usually be estimated by considering the number of detections, d , divided by the number of opportunities, n , present for detecting the cracks. In a designed experiment for estimating POD, the opportunities present at the time of inspection are known and only d is a random variable. However, the case of using flaws detected in the field, combined with knowledge of flaw growth and previous inspection times to infer missed opportunities, makes the “ n ” a random variable that is less than the true number of opportunities by the number of cracks that are undetected.

The effect of undetected cracks on a POD estimate is illustrated with the following construct. Consider that inspections are performed on a population of N cracks for k intervals. The cracks are postulated to be identical in growth characteristics so that at each inspection interval the same probability of detection applies across the crack population. Let p_i , $i = 1, 2, \dots, k$ be the probabilities of detection for each of the inspection periods. On average there will be Np_1 detected cracks in the first period, and at that time there is no knowledge of how many undetected cracks there are. However, of the $N(1-p_1)$ cracks that are undetected in the first period, $N(1-p_1)p_2$ will be detected in the second period (on average). These will generate a like number of misses that can be applied back to the first interval. Thus, after the second inspection the number of opportunities estimated for the first inspection interval would, on average, be $Np_1 + N(1-p_1)p_2$. If an estimate of the POD at time 1 is made after only two intervals of inspection, the estimate (using expectations) would be:

$$\hat{p}_1 = \frac{Np_1}{Np_1 + N(1-p_1)p_2} = \frac{p_1}{p_1 + (1-p_1)p_2} \quad (6.1)$$

Let $\hat{p}_{i,j}$ be the estimate made for probability of detection for the i^{th} inspection interval after j ($j > i$) inspections have been performed. Using the expected outcomes as above, it can be shown that the general estimate is given by:

$$\hat{p}_{i,j} = \frac{p_i}{1 - \prod_{l=1}^j (1 - p_l)} \quad (6.2)$$

Note that the denominator of the estimate is necessarily less than 1 and is the probability that a crack will have been detected by the j^{th} inspection.

Table 6-1 and Table 6-2 illustrate the effect of the missed cracks under two different scenarios. The first scenario is one in which there are 5 inspection periods and the crack growth is not very great so that the probability of detection does not grow very rapidly. The second scenario is one in which the crack growth is

fairly rapid so that the probability of detection is fairly high by the 5th inspection cycle. In both tables, the last column gives the proportion of cracks that have been found at some time in the 5 cycles of inspection.

Table 6-1: Slow Crack Growth or Slowly Increasing POD

Inspection cycle	1	2	3	4	5	
POD for cycle	0.1	0.12	0.15	0.2	0.25	Aggregate POD
Estimates at cycle 2	0.481					0.208
Estimates at cycle 3	0.306	0.367				0.327
Estimates at cycle 4	0.217	0.260	0.325			0.461
Estimates at cycle 5	0.168	0.201	0.252	0.336		0.596

Table 6-2: Fast Crack Growth or Sharply Increasing POD

Inspection cycle	1	2	3	4	5	
POD for cycle	0.09	0.26	0.4	0.5	0.75	Aggregate POD
Estimates at cycle 2	0.276					0.327
Estimates at cycle 3	0.151	0.436				0.596
Estimates at cycle 4	0.113	0.326	0.501			0.798
Estimates at cycle 5	0.095	0.274	0.421	0.527		0.949

In the case illustrated in Table 6-2, by the end of the 5th inspection cycle, 95% of the cracks have been found, and therefore the total number of cracks (detects and misses) included in the POD estimate for cycle 1 is close to the actual number of inspection opportunities.

In the above scenario, the quantification of the effect of missed cracks is straightforward. The POD estimates available for any given cycle are non-conservative by the factor $1/(1 - \text{prob of being undetected})$. The same effect will be present under more general crack growth models and when the POD function is written in terms of crack size. Simulations were used by Forsyth (2002) to demonstrate the non-conservatism under different conditions. It was concluded that the degree of non-conservatism is affected by discontinuity size distributions, inspection intervals and the steepness of the underlying POD; and without accurate knowledge of the underlying POD or of the actual discontinuity size distribution, it is impossible to determine the bias of a field data-based POD estimation. This work is described in detail in Annex I.

6.2 EFFECT OF CRACK SIZE AND SAMPLE SIZE ON POD

Previous studies have shown that the number and sizes of the cracks used in a POD capability evaluation have an interactive effect on the precision of the estimate, Berens and Hovey (1985). Since very little information is obtained from inspections of cracks of a size that are almost always detected or almost always missed, a disproportionate number of such cracks do not increase the precision of the POD estimate. In POD

evaluations using fabricated specimens, the number of cracks and the range of crack sizes can be controlled. However, with in-service inspection data, the number and sizes of available cracks are not controlled. Rather, both are determined by the real cracks detected by the inspections. The following summarize the results of studies that were performed to obtain an indication of the number of in-service cracks needed for a reasonably precise estimate of POD. Details of the studies are presented in Annex E.

6.2.1 POD(a) Model Approach

Simulated inspections were used to evaluate the joint effects of the number and sizes of cracks in a POD evaluation. The simulation process is described in Annex E. A log-normal POD(a) function with a 50% detectable crack size (a_{50}) of 50 mil and a 90% detectable crack size (a_{90}) of 190 mil was assumed. Sample sizes of 100, 300 and 500 cracks were drawn from crack size distributions of small, medium and large cracks. The relative sizes of the crack size distributions are defined with respect to the POD(a) function. The median crack sizes for the small, medium and large crack size distributions were 50, 100 and 150 mil, respectively. Fifty inspections were simulated for each of the nine combinations of sample and crack sizes. a_{90} and $a_{90/95}$ (the 95% confidence bound on a_{90}) were calculated for each of the inspections. The distributions of the estimates of these common POD characteristic values for the nine combinations provided the basis for comparing the combinations of sample size and crack size in POD evaluations. The results of these simulations are presented in Annex E.

To obtain an idea of the effect of the wrong POD(a) model, simulations were run with a cumulative log-normal distribution fit to data from true cumulative Weibull POD(a) models. Two Weibull models were used to determine whether or not the simulated inspections resulted in a hit or a miss. One of the Weibull models agreed with a cumulative log-normal equation at the a_{50} and a_{90} values. The second Weibull agreed with the cumulative log-normal at a_{90} , but with an a_{50} value half that of the cumulative log-normal. The results of these simulations are presented in Annex E.

When the cracks being used in a POD evaluation are small with respect to the a_{90} value of the POD(a) function, the results of the simulations clearly indicated the need for a large sample size. This conclusion is not at all surprising. When the cracks are all smaller than the a_{90} value, the estimate of a_{90} is obtained from an extrapolation. Although little is known about the population of crack sizes in the real structures being inspected, it is reasonable to assume that relatively few positive crack indications are obtained when compared to the total number of inspection sites in the entire fleet. That is, the sizes of the cracks in the structural detail are small in comparison to the target a_{90} value of the inspection system. Based on this assumption, it was concluded that the total number of inspection results (hits and misses) should be at least 300 in order to obtain a reasonably precise estimate of a_{90} . Even with 300 cracks in the analysis, significantly biased a_{90} values could be obtained if the true POD(a) is not reasonably modeled by the cumulative log-normal distribution.

Note that the use of in-service inspection data is a quite different scenario from that in which cracked specimens are used in a POD evaluation. The sizes of the cracks in specimens can be controlled, and are deliberately fabricated to cover the crack range of interest (see for example, Safizadeh et al. (2002)). When the crack sizes are relatively large with respect to the POD capability, fewer cracks are required and the results are not as sensitive to the wrong model. The data will be in the crack size range of interest.

6.2.2 CDF of Detected Cracks

The cracks detected at an inspection depend on both the sizes of the cracks that are in the structures and the POD capability of the inspection system. Given a probability density function for the crack sizes and a

POD(a) function for the inspection capability, the distribution of the sizes of the cracks that are expected to be detected can be calculated. A small study was performed to investigate the sensitivity of the CDF of detected cracks to various crack size distributions for a fixed POD(a) function. Detailed results of the study are presented in Annex F.

The results of this analytical study indicate that the cumulative distribution of the detected cracks depends jointly on the sizes of the cracks being inspected and the POD(a) function. If all cracks are small with respect to the a_{90} value of the inspection system, only small cracks will be detected and the 90th percentile of the detected cracks will be less than the a_{90} value. However, if the POD(a) is steep and very small cracks are not detected, the 90th percentile of the detected cracks can exceed the a_{90} value of the POD(a) function. At the time of an inspection, neither the sizes of the cracks in the structures nor the capability of the inspection system are known. Thus, the CDF of detected crack sizes is an uncertain estimate of the POD(a) capability of the inspection system. However, it is recognized that the CDF of the detected crack sizes provides an indication of the condition of the structure and verification that cracks of interest are being detected.

6.3 EFFECTS OF CRACK SIZE ERRORS/VARIABILITY ON POD ESTIMATION

In general, POD curves are estimated from inspections performed on cracks with assumed known lengths. If there are errors in the crack lengths, this will lead to error in the POD estimation. There are several ways that crack size error will manifest itself. A source of error is introduced in the back-calculation of a flaw size where an “average” crack growth rate is used, whereas it is known that, due to material variability, fatigue cracks of nominally identical sizes subjected to identical stress sequences will exhibit variability in size as a function of experienced stress cycles. Added to the back-calculation error is the possibility that the field measurement of crack may also be in error. This error can be a systematic reporting error or it may be a random error. The impacts of these two sources of crack size error are discussed here.

6.3.1 Inherent Crack Growth Variability

Fatigue cracks of nominally identical sizes subjected to identical stress sequences will exhibit variability in size as a function of experienced stress cycles. This inherent variability represents the absolute minimum scatter that will be present when back-calculating cracks sizes from detected cracks. To obtain some concept of the magnitude of this minimum degree of scatter over various back-calculation intervals, actual crack growth data from 68 exactly replicated tests on 2024-T3 aluminum panels were analyzed (Virkler, Hillberry and Goel (1978) or (1979) for the genesis of the data). The 68 histories of crack size versus number of cycles were translated to pass through a common crack size at three different cyclic lives. The scatter in the crack sizes at previous points in the history of the specimens are an indication of the lower bound of inherent variability that would be present regardless of the analytical crack prediction capability. Details of this investigation are presented in Annex G.

As an example, Figure 6-1 presents the inherent scatter when the recorded time histories are translated to pass through the original average crack size of 30mm at 216,000 cycles. The scatter at earlier numbers of cycles is representative of the inherent scatter that would be present given that the 30 mm crack had been detected at 216,000 cycles. The mean and standard deviation of back-determined crack sizes were calculated at 50,000 cycle intervals. The data indicated that a coefficient of variation of 5% would be a reasonable description of the inherent scatter for this material system and stress history. This degree of scatter can be introduced into the field measurement variability analysis of Section 6.3.2. An indication of the effect of this error combined with measurement error is given in the following section.

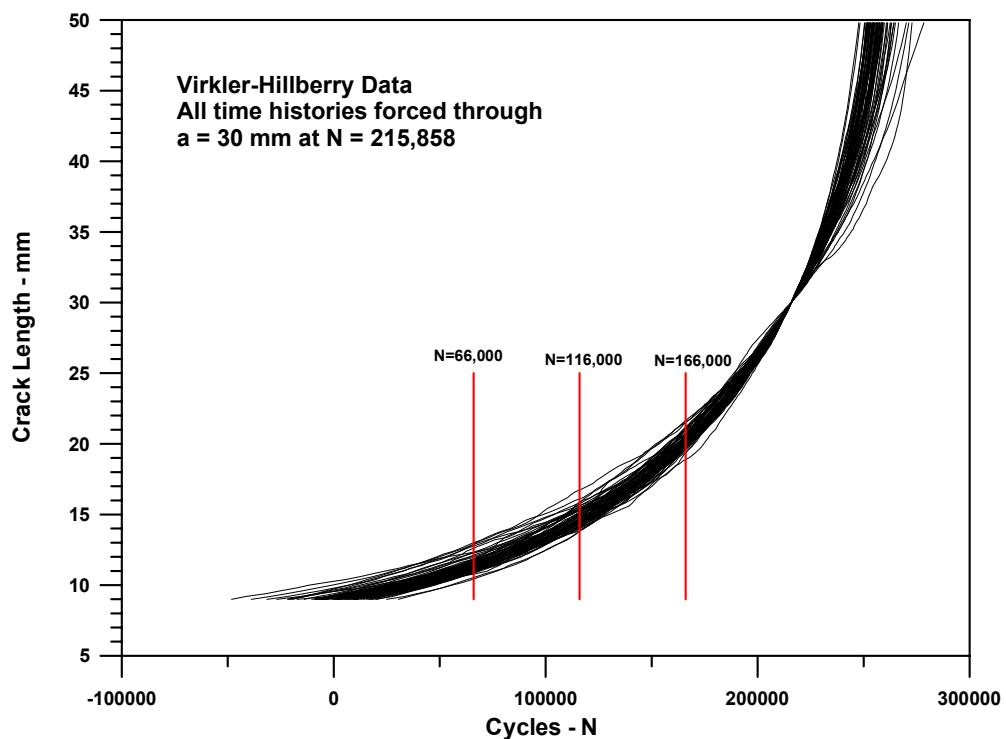


Figure 6-1: Actual Crack Growth Histories Forced to be Coincident at 30 mm.

6.3.2 Field Measurement Variability

Measurements made in the field are subject to error. However, the crack lengths used in calculating a POD curve are those that have been measured in the field (the hits) plus those crack lengths that have been inferred from back-calculations using an average crack growth curve. In this case, the hits and misses have different error sources for the crack lengths. The hits are subject to the field measurement errors, but the crack lengths for the misses are subject to: 1) the measurement error in the crack from which the miss lengths are inferred; 2) the possible error in the choice of appropriate crack growth mean line; and 3) the natural flaw variation discussed in the previous section. It is, therefore, likely that the overall variation in the flaw lengths used in POD estimation associated with the misses will be greater than that associated with the hits.

The impact of the sources of error are developed in Annex G from considerations of the underlying regression models that are used to estimate the parameters of the POD function as given in Section 5-2. If the parameter “ b ” is a bias in the crack length terms that applies across all the crack lengths used in the POD estimation, then the estimated mean for the POD function would have an expected value shifted from the true mean, that is, $E[\hat{\mu}] = \mu + b$. A random measurement error, “ σ_{δ}^2 ”, that applies across all the crack lengths affects the estimate of the variance parameter for the POD function. The effect is also influenced by whether there is a correlation of the error, “ $\sigma_{\varepsilon\delta}$ ”, with the actual flaw size. Specifically, the random error attenuates the estimate of the variance (with an adjustment due to correlation), that is $E[\hat{\sigma}^2] = \sigma^2 + \sigma_{\delta}^2 - 2 \cdot k \cdot \sigma_{\varepsilon\delta}$, where the constant, k , depends on the NDI process. If the crack length random error is independent of the length, the last term for the expectation of the variance estimate is 0.

SENSITIVITY OF POD TO IN-SERVICE INSPECTION DATA

The adjustments given above are in the logarithm – scale of the crack length measurements. They do provide a means to bound the influence on the estimated POD curves. To illustrate, consider the example given in Section 5.4.2 (details in Annex D) with the estimated POD function given in Figure 5-3.

In this example, the smallest crack size found was 0.019. Assuming that cracks could be measured to the nearest ± 0.005 , the relative error would be $\pm 26\%$. To derive a random error, this value is equated to 2 standard deviations so that $\sigma_\delta^2 \approx (0.13)^2 \approx 0.017$.

The total variation in crack lengths from measurement error plus the crack-to-crack differences discussed in Section 6.3.1 is $0.017 + 0.05^2 \approx 0.020$. However, the estimate of (μ, σ^2) from the example presented in Section 5.4.2 is $(-3.163, 0.688)$. Therefore, it is estimated that the measurement error and crack-to-crack material variation could have contributed up to 0.020 to this estimate of the variance, and the impact can therefore be assessed by considering the POD curve estimated by the parameters $(-3.163, 0.668)$. The difference in the crack length for which there is a 90% probability of detection is 0.002 inches. Comparing this to the 95% confidence bound shown in Figure 5-3 indicates that the random measurement error and the crack-to-crack variation (estimated at 5% relative error) are not significant contributors to the overall assessment of uncertainty for this example.

Chapter 7 – USE OF EXISTING POD INFORMATION

One of the major difficulties in using non-destructive inspection (NDI) data from in-service experience to determine NDI reliability is the typically small number of flawed sites. A set of in-service data provided by Heida (2001) contained 39 hits and 51 misses. No data on false call rates was available, as is typical for in-service inspection data. A set of data from a full-scale test at the Institute for Aerospace Research of the National Research Council Canada contained only 5 hits and 15 misses, and the inspection reliability could not be determined using traditional POD analyses (Forsyth et al. (2000)). In response to this common difficulty, pooling of data from similar NDI systems and structures has been suggested as a mechanism for increasing the size of data sets and thereby making traditional statistical analysis methods viable. Given the non-conservative bias inherent in the estimation of POD from field inspection data, estimation of POD from pooling is not recommended; the bias is not removed by the analysis of a larger set of biased data.

The key benefit desired from pooling data sets is to make use of existing data. Many POD experiments have been performed and reported in the open literature, and in cases where the existing data is applicable, one can theoretically at least avoid a duplication of effort and expense. The question on when data can be pooled is essentially the same as the question on whether one can use an existing POD – is this data set similar enough to my data? This chapter will provide a set of guidelines for the engineer to determine when he or she can use existing POD data alone or pooled in a new application. In discussions about NDI reliability or POD, there is a tendency to combine data sets or make extrapolations that may not be justified, as illustrated by the common question “what is the POD of eddy current (or x-ray or etc.)?”

Data pooling or using existing POD data for a new application requires the same information and fidelity as that in data collection and reporting, as described in Chapter 3 of this report. In addition, matching both the inspection procedures and the fidelity of data sets are required. Data requirements for POD outputs are the same or similar control and precision in the documented:

- known crack/artifact sizes,
- rigid “calibration” control, and
- rigid inspection procedure control including similarity in “acceptance criteria”.

On the simplest level, data pooling can be thought of as analogous to averaging, and many of the obvious advantages and pitfalls of averaging do apply. Averaging the heights of a sample of basketball players with the heights of a sample of football players will yield a result that tells us little about either basketball players or football players. Averaging the heights of samples of basketball players from different teams still yields useful information about basketball players.

The typical *raison d'être* for pooling is to increase the number of data points in order to apply statistical analysis methods to determine POD. Some guidelines exist for the number of data samples required in order to perform the common methods of estimating NDI reliability. The United States Department of Defense Handbook, MIL-HDBK-1823, suggests a minimum of forty flawed sites for “a vs. a” type systems. MIL-HBK-1823 also recommends that there be three times as many unflawed sites as flawed sites in order to estimate false call rates. Spencer et al. (1993) affirm this guideline, with the caveat that 30 flawed sites distributed between the 10th and 90th percentile of detectability is generally sufficient. An experimental corroboration of these guidelines is given in Forsyth et al. (2000). The distribution of the sizes of flaws in the data set is also important (see Safizadeh et al. (2002)), but cannot be controlled in the use of in-service data.

USE OF EXISTING POD INFORMATION

Data sets typical of in-service inspections may not be optimally distributed, and therefore higher numbers of flawed sites may be required.

The following sections describe some engineering and statistical checks that can be applied to NDI data to determine the potential for either pooling or simply using an existing POD in a new situation.

7.1 CRITICAL PARAMETERS FOR DATA SETS

From a simple mathematical viewpoint, two different data sets have the same POD and can be combined if they have the same “ \hat{a} vs. a ” response. From an engineering standpoint, combining results from dissimilar techniques or structures has little value, as the goal is to make an estimate of the reliability of similar inspections being performed on similar structures. The same ideas apply to using existing POD data in a new situation.

The method described in Annex J provides a simple test for data sets having the same of similar “ \hat{a} vs. a ” responses. This requires that multiple data sets have similar “calibration” artifacts and similar inspection procedures. Detail of the inspection procedures using the guidelines in Chapter 3 and Annex A may be validated by resultant similarity in “ \hat{a} vs. a ” responses.

Chapter 3 of this document describes the optimum data collection process to be followed in order to estimate NDI reliability from in-service data. Given two or more sets of data that have been collected with associated descriptive information described in Chapter 3, there are a number of critical parameters which must be matched in order for data pooling to provide a useful result. A simple “top-down” or flowchart approach can be applied to assess the potential usefulness of pooling different data sets, as outlined in the following text.

First, one can consider the specific NDI methods and techniques in question. There is little or no benefit pooling data from different NDI methods (for example eddy current and ultrasonics). If the desired result from pooling is an estimate of the POD for a particular technique applied to a particular type of structure, averaging over different NDI methods adds no information. More specifically, similarities in probe and technique must be examined. For example, a shear wave injected into a specimen at an angle will likely have different POD for a particular flaw type than a longitudinal wave at normal incidence. “Pencil-probe” eddy current inspections, even with different probes and instruments, may vary little in the resulting POD if the same calibration standard was used. This is the typical approach taken in commercial aircraft operation, where the maintainer is allowed to use any of a number of different equipment, provided they can achieve the same response on a calibration specimen. (Note: Calibration using a single artifact is not sufficient. Similarity in response to three or more artifacts that bound the desired detection capability are required to estimate POD from “ \hat{a} vs. a ” responses – see Annex J.)

In order to facilitate this process, an example flowchart has been developed. This is not meant to be a complete consideration of all possible cases, but to provide a guide allowing NDI practitioners to apply a logical decision process when faced with the question of pooling data sets.

As an example, assume that data exists for two similar situations. With aircraft type A, inspections are being performed using a high-frequency eddy current instrument and a pencil-probe. The object is to find surface-breaking fatigue cracks in aluminum 2024-T3 sheet material that is 1.5 mm thick. In aircraft type B, a different instrument and probe are being used also to find surface breaking cracks in 2024-T3 sheet material that is 2.5 mm thick. The flowchart should then assist the user in asking the key questions which will

determine whether or not these inspections are similar enough to have a “common” POD. An example is presented in the following text.

START:

- Inspection type?
 - eddy current (go to 1)
 - ultrasonic
 - ...
- 1 Eddy current specific techniques
 - high-frequency surface scan (go to 1.1)
 - bolt hole inspection (go to 1.2)
 - ...
- 1.1 Specimen material
 - Aluminum alloys (go to 1.1.1)
 - Ti alloys
 - Ferrous
 - Non-ferrous
 - ...
- 1.1.1 Material details
 - Al 2024T3 clad sheet (go to 1.1.1.1)
 - ...
- 1.1.1.1 Flaw characteristics
 - little or no residual stress (go to 1.1.1.1.1)
 - tightly closed
 - chemically affected crack surfaces (i.e. engine rotating components)
 - ...
- 1.1.1.1.1 Thickness
 - both greater than skin depth (go to 1.1.1.1.1.1)
 - ...
- 1.1.1.1.1.1 Probe diameter
 - both have same diameter within ?% (go to 1.1.1.1.1.1.1)
 - ...
- 1.1.1.1.1.1.1 Calibration
 - same calibration standard and screen responses (go to X)
 - ...

X: These inspection data can be pooled with high confidence in the applicability of the resulting POD to each inspection process.

7.1.1 Scaling “ \hat{a} vs. a ” Response to Pool Data

Much maintenance data does not contain information on the actual crack sizes detected, but simply reject when the response exceeds a set acceptance level. When POD capabilities are desired from maintenance data, precision measurement and recording of detected crack sizes is required. In the absence of crack size measurement information, an alternate method is described in Annex J to **estimate** POD capability using the “calibration” artifacts, representative cracks and specific inspection procedures as the basis for additional measurements and analyses. Fidelity of the method depends on cracks and “calibration” artifacts that are

representative of the population and rigid control of “calibration”, and the inspection procedure used for measurement.

7.2 ANALYSIS CHECKS ON DATA COMPATIBILITY

In concert with engineering criteria for data pooling as described in the previous section, there are statistical tests that can be applied to the question of whether multiple sets of data should be grouped together and characterized with a single POD curve. Consider the case where there are n identifiable populations of data specific to a non-destructive inspection. The n populations could be different operators, different periods of time in which data were taken, different probes, etc. The basic premise to be tested statistically is that the data in the individual populations can be considered to be outcomes from the same probability of detection curve.

Two of the most common ways to test statistically whether data across different conditions should be pooled are presented here. It should be emphasized that the tests presented here can only indicate that the existing data for different populations have different characteristics. However, it is possible that these differences are due to things other than different underlying POD curves. For example, consider pooling together the data from two different inspection programs for a single POD curve estimation. Also consider the two programs as truly having the same underlying POD in their inspections. However, the two programs are at different stages and have different numbers of missed cracks, as discussed in 7.2.3. The estimated POD curves from each of the populations will be significantly different, but for reasons other than true differences in the POD curves. The example in Section 7.2.3 shows fits to data that could well be explained by this phenomenon.

7.2.1 Likelihood Ratio Tests

Given a POD function, $\pi(a)$, where a is the crack length, the outcome of a detection is said to have likelihood of $\pi(a)$, and the outcome of a miss is said to have likelihood of $1 - \pi(a)$, that is, the probability of a miss. The likelihood associated with a set of inspection outcomes is the product of the individual crack likelihoods. Letting i index the detected cracks and letting j index the missed cracks, the likelihood, L , can be written as

$$L(\pi) = \prod_i \pi(a_i) \cdot \prod_j (1 - \pi(a_j)) \quad (7.1)$$

The general statistical procedure that is used to estimate the POD function is to find the set of parameters for the function $\pi(a)$ that maximizes the total likelihood given in equation (7.1). However, it is easier mathematically to work with logarithms of likelihood and therefore the log-likelihood is defined as

$$LL(\pi) = \ln(L(\pi)) = \sum_i \ln(\pi(a_i)) + \sum_j \ln(1 - \pi(a_j)), \quad (7.2)$$

where, as before, i indexes the detected cracks and j indexes the missed cracks. Considering the ratio of their likelihoods, or equivalently the difference in their log-likelihoods, compares two different POD functions that could be used to explain the data.

Consider the case of having n individual populations of “hit/miss” data. The POD function given by equation (5.1) in Section 5.2 is fit to each population. Let $\pi_i(a), i = 1, 2, \dots, n$ be the individual fits. The log-likelihood for all the data is the sum of the individual log-likelihoods and is given by the following formula:

$$LL_T(\pi_1, \pi_2, \dots, \pi_n) = \sum_{i=1}^n \ln(L_i(\pi_i)). \quad (7.3)$$

It should be emphasized that each of the likelihood functions indexed by i in equation (7.3) are restricted to the data of population i . Each of the population POD functions, $\pi_i(a)$, are defined by two parameters, where those two parameters were chosen to maximize the likelihood (and the log-likelihood, since the logarithm is a monotonic function). Therefore, a total of $2n$ parameters are used to characterize the total data set. If the data are pooled, a single POD function would be used to characterize the data using only two parameters and the log-likelihood of the pooled data is given by

$$LL_P(\pi_0, \pi_0, \dots, \pi_0) = \sum_{i=1}^n \ln(L_i(\pi_0)). \quad (7.4)$$

Statistical theory tells us that the quantity, G , defined by

$$G = 2 \cdot [(LL_T(\pi_1, \pi_2, \dots, \pi_n) - LL_P(\pi_0, \pi_0, \dots, \pi_0))] = \sum_{i=1}^n 2 \cdot [\ln(L_i(\pi_i)) - \ln(L_i(\pi_0))]. \quad (7.5)$$

has a distribution that is chi-square with $2n - 2$ degrees of freedom under the hypothesis that the n populations are all governed by a single 2-parameter POD function. The log-likelihoods as defined above are easily calculated for given POD functions and are usually part of the available output in commercial software used to fit binary regression data. Larger G -values indicate that the added parameters are significant and that the data should not be pooled. Letting the symbol $\chi^2(\nu)$ denote a chi-square random variable with ν degrees of freedom, the decision to pool the data from all the populations can be made from the associated p -value given by $p = \Pr(\chi^2(2n - 2) > G)$. The p -value is the probability that the variation in the fitted POD curves would have occurred by random, when the various populations were actually governed by a single POD curve. If the probability is low, then differences in POD curves are taken to be the likely reason for the high G -value. It is common to use a level of 0.05 as the decision level and thus pool the data if $p > 0.05$.

If multiple populations are being considered at the same time, the above procedure could indicate that the populations should not be pooled, when it is only one of the populations that should not be pooled with the rest. This can be judged by looking at the individual terms in the summand on the right-hand side of equation (7.5). If a single population is a large contributor to the difference, it can be removed, and the above procedure followed to decide on the pooling of the remaining $n - 1$ populations.

NOTE: The theory behind the above procedure for pooling data is based on the existence of model parameterizations such that an added parameter being zero is equivalent to dropping the parameter in the model. Hosmer and Lemeshow (1989) give a general discussion of the use of likelihood ratio tests for binary data.

7.2.2 Wald (Coefficient Standard Error) Tests

A useful property of maximum likelihood estimators is that they are asymptotically distributed as Normal random variables as the sample size increases. Therefore, the ratio of the estimate to its standard error can be judged against the standard Normal distribution to determine significance. This is the method that is often used in software packages to indicate the level of significance. It requires the estimation of the standard error for the parameters of the model. These estimates come from the matrix of second partial derivatives of the log-likelihood function taken with respect to the model parameters.

This method for comparing multiple POD curves (or populations) is treated in more detail in Annex H of MIL-HDBK-1823 (1999).

7.2.3 Example Comparison

To illustrate the methodology suggested for checking the compatibility of data pooling across populations, the example from Section 5.2.2 will be considered. In that example, 39 detected cracks were combined with 51 misses that were calculated from back-growth models using crack growth severity stress calculated by aircraft. The data are listed in Table D-3 of Annex D. Consider the question of whether the data from the aircrafts numbered 1 to 19 are consistent with that taken from aircrafts 20 to 39. This breakdown is arbitrary and for illustration purposes. There is no a-priori reason for suspecting a difference in the populations.

The split of the data leaves the two populations roughly equivalent, with the first population containing 19 detected cracks and 26 misses and the second population containing 20 detects and 25 misses. The original fit given in Section 5.2.2 is for the pooled data and results in a pooled log-likelihood of $LL_p = -54.12 = -26.12 + -28.00$, where partition of the log-likelihood is also shown. Fitting each population independently and combining the data results in a total log-likelihood given by $LL_T = -50.51 = -23.05 + -27.46$. Applying equation (7.5)

$$G = 2 \cdot [(LL_T(\pi_1, \pi_2) - LL_p(\pi_0, \pi_0))] = \sum_{i=1}^2 2 \cdot [\ln(L_i(\pi_i)) - \ln(L_i(\pi_0))] = 6.14 + 1.08 = 7.22.$$

The p -valued associated with 7.22 for a chi-square distribution with 2 degrees of freedom is 0.027, and there is evidence that the calculated POD functions for the individual populations are different and should not necessarily be pooled. Figure 7-1 shows the POD fits the pooled data as well as to each of the populations.

This example also illustrates a potential impact of the effect of undetected cracks as discussed in Section 6.1. That is, the POD curve fit to the first population is overly optimistic and will be substantially different the first time the data include a larger flaw that implies some misses at cracks larger than 0.041. To illustrate the sensitivity to this phenomenon, consider dividing the populations into aircraft numbers 1 – 6, 8 – 20 as the first population, and the second population being aircraft numbers 7 and 21 – 39. There is the same number of data points in these two populations as in the original two populations. The difference is the exchange of a detected crack at 0.03 and its inferred miss at 0.027 (AC 7), with a detected crack at 0.15 and its inferred miss at 0.076 (AC 20). With this one small change, $G = 1.22 + 0.66 = 1.88$, and the p -value is 0.39, implying that the differences in the two fit POD curves can be attributed to statistical variation and the data can be pooled from the two populations. The individual fit POD curves are also shown in Figure 7-1.

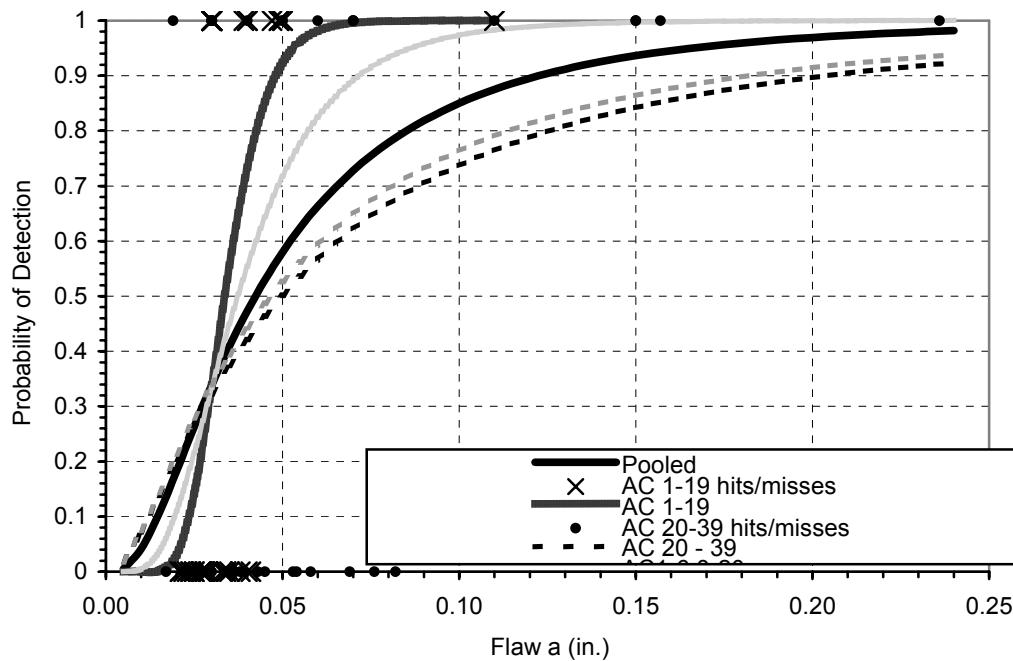


Figure 7-1: Comparison of 2 Populations of in-Service Inspection Data from F-16 ASIP Control Point.

7.3 CHECKS ON ASSUMED POD

Given a POD function, a flaw size population and loading and flaw growth information, one can determine the expected results of an inspection on a set of components. Although these requirements do sound idealistic, they are all required to use damage tolerance, retirement for cause, or safety-by-inspection maintenance philosophies. By comparing the results of actual inspections during the maintenance process with the expected results, it is possible to make an assessment of the validity of the original assumptions. However, it is not possible from this analysis to determine which of the ingredients is off, as any one of POD, flaw population, loading or flaw growth may change the results seen in an actual aircraft (or other) maintenance situation.

The find or miss decision for some advanced inspection systems is based on the magnitude of a numerical response signal, \hat{a} , where the decision threshold, \hat{a}_{dec} , is determined from an evaluation of the NDI response from specimens with known crack sizes, Petrin et al. (1993). For such inspections, a check on the consistency between field and demonstration inspections can be performed quite simply. The signal magnitudes and crack sizes can be compared with the “ \hat{a} vs. a ” of the POD demonstration for compatibility. To perform this check will require measuring and preserving the size of in-service detected crack and its corresponding NDI response, \hat{a} .

7.4 A BAYESIAN METHOD TO POOL DATA

A Bayesian-based method to use new data such as that from field inspections to modify an assumed POD curve was derived in Leemans and Forsyth (2004). A brief description of that method is provided in this section. Although this paper is written using the log-logistic function as the assumed model of the POD curve, the analysis also applies to other models such as the more commonly used log-normal.

USE OF EXISTING POD INFORMATION

The Bayesian framework will be used here to modify estimates of the parameters (α and β) of a log-logistic model for POD of the following form (Berens and Hovey, 1983):

$$P_i = \frac{\exp[\alpha + \beta \ln(a_i)]}{1 + \exp[\alpha + \beta \ln(a_i)]} \quad (7.6)$$

where P_i is the probability of detection of crack i , a_i is the length of crack i , and α and β are parameters of the model.

However, this is not sufficient to proceed with the evaluation of the effect of the new evidence on the estimated parameters of the model. If the model parameters (α and β) were perfectly known, then the field inspection data would not change the value of these parameters, and therefore the POD curve would not be changed by the presence of the evidence of the field inspections. What is further required to carry out the Bayesian analysis is to identify α and β as random variables whose distribution can be described by the joint prior distribution $P_{prior}(\alpha, \beta)$ of the parameters α and β . If this prior information were available, then Bayes' Theorem can be used to calculate the posterior probability of α and β based on field inspection data

$$P_{posterior}(\alpha, \beta | fielddata) \propto Likelihood(fielddata | \alpha, \beta) P_{prior}(\alpha, \beta) \quad (7.7)$$

where the likelihood of the field data is the product of the probabilities that the result of the inspection occurred as it did for each field inspection (independence of each inspection is assumed), given that the parameters of the log-logistic model are α and β .

$$Likelihood(fielddata | \alpha, \beta) = \prod_{i=1}^r probability(occurrence_i | \alpha, \beta) \quad (7.8)$$

The posterior probability density for α and β defined in equation (7.7) can then provide estimates of the means of the new parameters $\alpha_{posterior}$ and $\beta_{posterior}$ for an updated model of the POD.

Up to the present, only two options were available to the analyst: 1) ignore historical data and use only the small data set based on field inspections; and 2) or ignore the field inspection data and base future inspection scheduling on historical POD curves which may well not be based on inspections of the specific component. The approach developed in this paper demonstrates how the two types of information may be systematically combined.

It should be noted that the Bayesian framework developed in this report gives reasonable results. When the prior distribution on the POD curve is well defined and has little uncertainty, then new evidence from a few field inspection data hardly changes the POD curve. However, when the prior distribution on the POD curve is ill defined and has large uncertainties, then new evidence alters the POD curve in a meaningful way. The ranking from total knowledge (the prior is perfect) via narrow prior and wide prior to total ignorance makes very much sense.

The effect of the unknown and unavailable misses in field data has not been evaluated in this model.

7.5 SUMMARY

This chapter addressed two common needs: 1) data pooling in the case of “ \hat{a} vs. a ” type data is desirable to increase confidence level in POD outputs; and 2) in a situation where validated POD data is not available, it may be possible to use existing data. These problems are very similar in that they require an assessment of the similarity between two or more inspection situations. The same rigor in assessment of data quality must be applied to individual candidate data sets and the resultant pooled data. Requirements include:

- known crack/artifact sizes,
- rigid “calibration” control, and
- rigid inspection procedure control including similarity in “acceptance criteria”.

In addition, similarity of data sets is required and must be tested to ensure expectations of similarity in results. The method described in Annex J is suggested as a minimum requirement for both the testing of data sets and in documentation of combined/pooled data sets.

USE OF EXISTING POD INFORMATION



Chapter 8 – CONCLUSIONS

The intent of this Working Group was to evaluate the potential of reducing life cycle costs while ensuring flight safety through the use of real field inspection-based probability of detection data. As stated in the Introduction, most NDI reliability data available results from dedicated round-robin inspection programs whereby the same samples are inspected by disparate technicians under laboratory type, or in some cases, simulated in-service conditions. These data have been frequently challenged on the basis that the inspection conditions in terms of environment, access and human factors may not be representative of those seen in service. Analysis of in-service NDI findings can improve our understanding of the performance of NDI. This greater confidence in NDI reliability would allow more effective use of NDI for life extension.

The technical challenge addressed by this Working Group was to define processes to use the significant numbers of service detections to characterize NDI reliability through the calculation of a Probability of Detection. The general approach investigated was as follows: 1) for a detected crack, determine a characteristic measurement that can be used in a crack growth study; 2) based on detailed knowledge of the component and its usage history, ‘back-calculate’ the size of the crack to the initial detectable size; 3) using the calculated crack size history and the inspection history of the part, determine the size of the ‘missed’ cracks; and 4) using the detected size and the missed size data, calculate the field POD for the inspection technique.

The Working Group reached the following positive conclusions:

- Detailed processes for collecting, documenting and pooling in-service inspection results are both possible and feasible. Within a specific country, this could be done in a relatively straightforward manner using internal management procedures. To collate data from a number of countries would be complicated and would require a concerted effort to define and implement detailed procedures.
- Analytically, it is feasible to determine the ‘missed’ crack sizes using a combination of ‘back-growth’ calculations’ and information from inspections performed.

As part of the sensitivity studies performed late in the Working Group study, a fundamental flaw in using this derived information to determine the POD of the applied inspection technique was identified. The process defined above only uses the ‘missed’ data associated with detected cracks. Through detailed evaluations (Section 6) of the process, it was determined that statistically this process provides a non-conservative value of POD. Essentially, in addition to the ‘misses’ identified through the process of ‘back-growth’ for detected cracks, there is another population of ‘misses’ for cracks that have not yet been detected. This non-conservatism leads to the conclusion that “hit/miss” populations derived from detected cracks alone are insufficient to provide a useable POD for the technique.

The Working Group investigated other data reduction techniques of this data set that could provide important information on the inspection reliability in a field environment. Two in particular were evaluated. These were the Cumulative Distribution Function (CDF) and the Binomial Model/Bayesian Approach. The conclusions from these studies were:

- The CDF of detected crack sizes is an uncertain estimate of the POD(a) capability of the inspection system. However, it is recognized that the CDF of the detected crack sizes provides an indication of the condition of the structure and verification that cracks of interest are being detected. The CDF of detected crack sizes does provide information about the capability of the NDI system in the in-service environment. The curve is quite stable and less vulnerable to changes in the data set than the POD

CONCLUSIONS

curve. The CDF does not directly yield the reliably detectable crack size (at a given confidence level), but it gives a first estimate of this size.

- The use of Bayesian inference may be able to give estimates for safety levels on very limited data.

The Working Group membership consisted of NDI practitioners, statisticians, structural engineers, life cycle managers and regulators. As such, it was uniquely competent to address a broad range of topics associated with NDI reliability and its influence on life cycle management and airworthiness. Documents tabled as part of the on-going interactions offer a wealth of information on established procedures, and in many cases, are unique in supplying a concise explanation of how we arrived at current practice and they are included as Annexes.

Chapter 9 – RECOMMENDATIONS

The key conclusion of this report is that in-service inspection data analyzed simply using existing probability of detection methodology will produce a non-conservative estimate of POD. However, the body of work in this report and its Annexes include significant information which can assist the operator of a fleet in assessing the capability of an inspection, and therefore the risk associated with the inspection regime.

Increased attention must be paid to repeatability and reproducibility of inspections in order to achieve the maximum POD of in-service inspections. The use of relevant and traceable calibration artifacts, multi-point calibration and training on naturally cracked specimens are simple and effective means of improving inspection performance, but rarely implemented.

In many practical cases, when unexpected cracking or other damage is found in a fleet, decisions on inspections and risk must be made quickly without resorting to extended experimentation. The guidelines in this report for assessing the applicability of existing POD data for use in a new situation provide a means to support decision in this regard. Formal guidelines or regulations for procedures to be followed in this common situation should be developed, using the information in this report as a starting point.

The inspection findings in any fleet maintenance situation are a combination of the crack or discontinuity size distribution at the time of inspection and the POD of the inspection technique employed. Therefore, it may be possible to assess the validity of the assumed crack size distribution and POD from the in-service data. Further research is required to develop this potential use of in-service inspection data.

Alternate methods of estimating inspection performance using the cumulative distribution function or Bayesian methods have been proposed in this report. These methods should be investigated further in order to fully assess their capability.

Finally, it was found that very little in-service inspection data in any NATO country is being recorded with sufficient information in order to allow its use in further analysis of fleet cracking or inspection performance. Minor improvements to data recording can provide significantly more useful information on both the populations of cracks that exist in fleets as well as on the inspection performance than is currently available in most NATO countries.

RECOMMENDATIONS



Chapter 10 – GLOSSARY

For the purposes of this document, we make the following definitions.

a_{90} – crack size for which there is 90% probability of detection.

$a_{90/95}$ – upper 95% confidence limit on an estimate of a_{90} .

Back-calculation – using a fracture mechanics-based crack growth versus time relation to estimate crack sizes at earlier times.

Binomial POD analysis – the approach to characterizing inspection capability in which a constant detection probability is assumed for sample of inspection results being analyzed.

Bayesian binomial POD analysis – the analysis of binomial POD data in which the uncertainty in the estimate of POD is modeled by a prior distribution that is updated by data.

Cumulative Distribution Function (CDF) – a summary of data that expresses the proportion of a population that is less than the argument.

Cumulative log-normal model – a standard CDF of statistics that has been found acceptable as a model for the $POD(a)$ function.

Exceedance probability – expresses the proportion of a population that is greater than the argument, i.e. $1 - CDF$.

False call – a “false call” occurs when an inspection technique is applied to a location with no flaw of any size and the inspection technique indicates the existence of a flaw.

Hit – a “hit” occurs when an inspection technique is applied to a flawed location and the inspection technique indicates the existence of the flaw. The existence of the flaw must be verified.

Maximum likelihood – a parameter estimation method that maximizes the probability of obtaining a particular set of results.

Miss – a “miss” occurs when an inspection technique is applied to a flawed location and the inspection technique does not indicate the existence of the flaw, regardless of flaw size.

POD(a) – the proportion of all cracks of size a that will be detected by the NDI system when applied by representative inspectors to the population of structural elements in a defined environment.

POD(a) model – the approach to characterizing inspection capability in which a specific model is assumed for the $POD(a)$ function and the inspection results are used to estimate the parameters of the model.

μ – location parameter of the cumulative log-normal model. $\exp(\mu)$ is a_{50} , the 50% detectable crack size.

σ – scale parameter of the cumulative log-normal model. $\exp(\mu + 1.282 * \sigma)$ is a_{90} , the 90% detectable crack size.

GLOSSARY



Chapter 11 – REFERENCES

Asada, H., Sotozaki, T., Endoh, S. and Tomita, H. (1998), "Practical Evaluation of Crack Detection Capability for Visual Inspection in Japan," RTO-MP-10 AC/323(AVT)TP/2, RTO Meetings Proceedings 10, *Airframe Inspection Reliability under Field/Depot Conditions*, NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex, France, November, 1998.

Berens, A.P. and Hovey, P.W. (1981), *Evaluation of NDE Reliability Characterization*, AFWAL-TR-81-4160, Volume 1, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio, December, 1981.

Berens, A.P. and Hovey, P.W. (1982), "Characterization of NDE Reliability", *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 1, ed. D.O. Thompson and D.E. Chimenti, Plenum Press, 1982.

Berens, A.P. and Hovey, P.W. (1984), *Flaw Detection Reliability Criteria, Volume I – Methods and Results*, AFWAL-TR-84-4022, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio, April, 1984.

Berens, A.P. and Hovey, P.W. (1985), "The Sample Size and Flaw Size Effects in NDI Reliability Experiments," *Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 4B, ed. D.O. Thompson and D.E. Chimenti, Plenum Press, 1985, pp. 1327-1334.

Berens, A.P. (1988), "NDE Reliability Data Analysis," *ASM Metals Handbook, Volume 17, 9th Edition: Nondestructive Evaluation and Quality Control*, ASM International, Materials Park, Ohio, 1988, pp. 689-701.

Berens, A.P., West, J.D. and Trego, A. (1998), "Risk Assessment of Fatigue Cracks in Corroded Lap Joints," RTO-MP-18 AC/323(AVT)TP/8, RTO Meetings Proceedings 18, *Fatigue in the Presence of Corrosion*, NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex, France, November, 1998.

Brewer, J.C. and Mengert, P. (1992), "Preliminary Output of the Aircraft System Reliability Analysis," Project Memorandum, DOT-VNTSC-FA3H2-93-13, U.S. Department of Transportation, John A. Volpe National Transportation System Center, Cambridge, MA, September, 1992.

Brewer, J.C. (1993a), "Reliability Analysis Using Maintenance Data," in the *Proceedings of the First FAA Inspection Reliability Workshop*, Atlantic City, NJ, September, 1993.

Brewer, J.C. (1993b), "Simulation of Nominal Crack Growth Histories," Letter Report, LTR-FA3H2-9302, U.S. Department of Transportation, John A. Volpe National Transportation System Center, Cambridge, MA, September, 1993.

Brewer, J.C. (1993c), "Implications of Estimated Probability of Detection Curves," Technical Interchange, U.S. Department of Transportation, John A. Volpe National Transportation System Center, Cambridge, MA, December, 1993.

Brewer, J.C. (1994a), "Cumulative Distribution of Detected Cracks: Service Difficulty Report Data and Analytical Models," Letter Report LTR-FA4H2-9404, U.S. Department of Transportation, John A. Volpe National Transportation System Center, Cambridge, MA, April, 1994.

REFERENCES

Brewer, J.C. (1994b), *Estimate of Probability of Crack Detection from Service Difficulty Report Data*, DOT/FAA/CT-94/90, U.S. Department of Transportation, FAA Technical Center, Atlantic City International Airport, September, 1994.

Brewer, J.C. (1995), "Reliability Analysis using Japanese Maintenance Data," in the *Proceedings of the Second FAA Inspection Reliability Workshop*, Atlantic City, NJ, January, 1995.

Brewer, J.C., Mengert, P. and Disario, R. (1996), *Probability of Visual Crack Detection from Japanese Maintenance Data Using Survival Analysis*, Draft Report REP-FA6H2-9601, U.S. Department of Transportation, John A. Volpe National Transportation System Center, Cambridge, MA, January, 1996.

Bruce, D.A. (1998), "NDT Reliability Estimation from Small Samples and In-Service Experience," RTO-MP-10 AC/323(AVT)TP/2, RTO Meetings Proceedings 10, *Airframe Inspection Reliability under Field/Depot Conditions*, NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex, France, November, 1998.

Castner, W.L., Forman, R. and Rummel, W.D., Agreement and Private Communication.

Cochran, J.B., Bell, R.P., Alford, R.E. and Hammond, D.O. (1991), "C-141 WS405 Risk Assessment", in the *Proceedings of the 1991 USAF Structural Integrity Program Conference*, San Antonio, Texas.

Endoh, S., Tomita, H., Asada, H. and Sotozaki, T. (1993), "Practical Evaluation of Crack Detection Capability for Visual Inspection in Japan," *Proceedings of the 17th Symposium of the International Committee on Aeronautical Fatigue (ICAF)*, 1993, pp. 259-280.

Fahr, A., Forsyth, D.S. and Bullock, M. (1993), "A Comparison of Probability of Detection (POD) Data Determined using Different Statistical Methods", National Research Council Canada, Institute for Aerospace Research Report LTR-ST-1947, December, 1993.

Fahr, A., Forsyth, D.S., Bullock, M., Wallace, W., Ankara, A., Kompotiatis, L. and Goncalo, H.F.N. (1995), *POD Assessment of NDI Procedures using a Round-Robin Test*, AGARD-R-809, Advisory Group for Aerospace Research and Development, NATO, Neuilly-sur-Seine Cedex, France, January, 1995.

Forsyth, D.S. and Fahr, A. (1998), "An Evaluation of Probability of Detection Statistics," RTO-MP-10 AC/323(AVT)TP/2, RTO Meetings Proceedings 10, *Airframe Inspection Reliability under Field/Depot Conditions*, NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex, France, November, 1998.

Forsyth, D.S., Boate, K., McRae, K.I. and Fahr, A. (1999), *The Use of Data from the CT-114 Full Scale Test for the Development of POD*, National Research Council Canada, Institute for Aerospace Research Report LTR-SMPL-1999-0136, November, 1999.

Forsyth, D.S., Leemans, D.V., Fahr, A. and McRae, K.I. (2000), "Development of POD from In-Service NDI Data", in *Review of Progress in Quantitative Nondestructive Evaluation*, ed. D.O. Thompson and D.E. Chimenti, AIP Conference Proceedings 509, Vol. 19B, 2000, pp. 2167-2174.

Forsyth, D.S., Safisadeh, M.-S. and Fahr, A. (2002), "Issues in the Determination of Probability of Detection using Field Inspection Data", in the Proceedings of the 3rd European – American Workshop on the Reliability of NDE and Demining, Berlin, 10-13 September, 2002.

Harding, C.A. and Hugo, G.R. (2003), "Statistical Analysis of Probability of Detection Hit/Miss Data for Small Data Sets", in *Review of Quantitative Nondestructive Evaluation*, ed. D.O. Thompson and D.E. Chimenti, AIP Conference Proceedings 657, Vol. 22, pp. 1838-1845.

Harris, J.A. Jr. (1987), *Engine Component Retirement for Cause, Volume I – Executive Summary*, AFWAL-TR-87-4069, Air Force Wright Aeronautical Laboratories, Wright-Patterson Air Force Base, Ohio 45433-6533.

Heida, J.H. and Grooteman, F.P. (1998), "Airframe Inspection Reliability using Field Inspection Data", RTO-MP-10 AC/323(AVT)TP/2, RTO Meetings Proceedings 10, *Airframe Inspection Reliability under Field/Depot Conditions*, NATO Research and Technology Organisation, Neuilly-sur-Seine Cedex, France, November, 1998.

Heida, J.H. (2001), Presentation to AVT-051 Committee Meeting, Loen, Norway, 2001.

Hosmer, D.W. Jr. and Lemeshow, S. (1989), *Applied Logistic Regression*, Wiley, New York, 1989.

Koul, A.K., Thamburaj, R., Wallace, W., Raizenne, M.D. and de Malherbe, M.C. (1985), "Practical Experience with Damage Tolerance Based Life Extension Concepts for Turbine Engine Components", in the proceedings of the AGARD-SMP *Conference on Damage Tolerance Concepts for Critical Engine Components*, AGARD-CP-393, San Antonio, Texas, 1985, pp. 23-1 to 23-22.

Leemans, D.V. (1998), "Probability of Detection Based on Field Inspection Data," Final Project Report, Institute for Aerospace Research of the National Research Council Canada, December, 1998.

Leemans, D.V. (2000), "Probability of Detection Based on Field Inspection Data: A Practical Application," Project Report, Air Vehicles Research Section, Department of National Defence, Canada, November, 2000.

Leemans, D.V. and Forsyth, D.S. (2004), "Bayesian Approaches to Using Field Inspection Data in Determining the Probability of Detection", Accepted for Publication in Materials Evaluation.

Lewis, W.H., Sproat, W.H., Dodd, B.D. and Hamilton, J.M. (1978), *Reliability of Nondestructive Inspections – Final Report*, SAALC/Mme 76-6-38-1, San Antonio Air Logistics Center, Kelly Air Force Base, TX 78241, December, 1978.

Lincoln, J.W. (1997), "Risks Assessments of Aging Aircraft," *Proceedings of the First Joint DOD/FAA/NASA Conference on Aging Aircraft*, Ogden, Utah.

Lincoln, J.W. (2000), "Aging Systems and Sustainment Technology," Structures Technology for Future Aerospace Systems, Ahmed K. Noor, Editor, Volume 188, Progress in Astronautics and Aeronautics.

Lockheed Martin Corporation (1997), "Nondestructive Inspection; USAF/EPAF Series F-16A and F-16B Aircraft", Technical Manual T.O. 1F-16A-36, Change 38, 3 March, 1997.

MIL-HDBK-1823 (1999), *Nondestructive Evaluation System Reliability Assessment*, United States Department of Defense Handbook, MIL-HDBK-1823, 30 April, 1999.

Miller, M. (1995), "Visual Detection Reliability During Routine Maintenance," *Proceedings of the Second FAA Inspection Reliability Workshop*, Atlantic City, NJ, January, 1995.

REFERENCES

Petrin, C., Annis, C. and Vukelich, S.I. (1993), *A Recommended Methodology for Quantifying NDE/NDI Based on Aircraft Engine Experience*, AGARD-LS-190, Advisory Group for Aerospace Research and Development, NATO, Neuilly-sur-Seine, France, April, 1993.

Roth, P.G. (1992), *Probabilistic Rotor Design System (PRDS) Phase I*, WL-TR-92-2011, Wright Laboratory, Wright-Patterson Air Force Base, Ohio 45433, May, 1992.

Rummel, W.D., Todd, P.H. Jr. and Castner, W.L. (1973), "Detection of Fatigue Cracks by Nondestructive Evaluation Methods", Paper presented at the Spring Convention, American Society for Nondestructive Testing, Los Angeles, California, March, 1973.

Rummel, W.D., Frecska, S.A. and Rathke, R.A. (1974), *The Detection of Fatigue Cracks by Nondestructive Testing Methods*, NASA CR-2369, February, 1974.

Rummel, W.D. (1982), "Recommended Practice for Demonstration of Nondestructive Evaluation (NDE) Reliability on Aircraft Production Parts", Materials Evaluation, Vol. 40, No. 9, August, 1982, pp. 923-932.

Safizadeh, M-S., Forsyth, D.S. and Fahr, A. (2002), "Recent Studies on the POD Analysis of "a" vs. "a" NDI Data", in *Review of Progress in Quantitative Nondestructive Evaluation*, ed. D.O. Thompson and D.E. Chimenti, AIP Conference Proceedings Volume 657, Vol. 22, 2002, pp. 1846-1853.

Schütz, W.H. (1996), "A History of Fatigue", Engineering Fracture Mechanics, Vol. 54, No. 2, 1996, pp. 263-270.

Simpson, D.L. (1981), *Development of Non-Destructive Inspection Probability of Detection Curves using Field Data*, National Research Council Canada, Institute for Aerospace Research Report LTR-ST-1285, August, 1981.

Spencer, F.W., Borgonovi, G., Roach, D., Schurman, D. and Smith, R. (1993), *Reliability Assessment at Airline Inspection Facilities Volume I: A Generic Protocol for Inspection Reliability Experiments*, DOT/FAA/CT-92/12, I, March, 1993.

Virkler, D.A., Hillberry, B.M. and Goel, P.K. (1978), *The Statistical Nature of Fatigue Crack Propagation*, AFFDL-TR-78-43, Air Force Flight Dynamics Laboratory, Wright-Patterson Air Force Base, Ohio, April, 1978.

Virkler, D.A., Hillberry, B.M. and Goel, P.K. (1979), "The Statistical Nature of Fatigue Crack Propagation," Journal of Engineering Materials and Technology, ASME, Vol. 101, April, 1979, pp. 148-152.

Yang, J.N. and Chen, S. (1985), "Fatigue Reliability of Gas Turbine Engine Components Under Scheduled Inspection Maintenance", Journal of Aircraft, Vol. 22, No. 5, May, 1985, pp. 415-422.

Annex A – DATA COLLECTION PROCESS

A.1 OVERVIEW

Data requirements for use in developing Probability of Detection (POD) outputs are:

- known crack/artefact sizes,
- rigid calibration control, and
- rigid procedure control.

The usefulness of maintenance data collected is dependent in large part on the fidelity and precision of that data. Non-destructive inspection (NDI) utilizes indirect measurement of a material characteristic or parameter and correlation of that measurement to a desired material characteristic or property. Reliable detection of cracks (or other discontinuities) by an applied (NDI) procedure is dependent on:

- capability,
- reproducibility, and
- repeatability.

The CAPABILITY of a procedure is roughly characterized by the inherent signal and noise responses as applied to a specific test object and crack-to-crack variances within the test object. The capability and hence applicability of an NDI procedure is dependent on the fidelity and precision of the causal model relationship between the measured parameters (NDI output) and the desired characteristic. This is inherent in the physics of the NDI method and application parameters including the threshold limit used for purposes of accept or reject.

The REPRODUCIBILITY of a procedure is generally characterized by the inherent capability and variances in the procedure “calibration” process. Reproducibility is defined as the ability for a specific NDI technique to be performed or “reproduced” from a set of specifications. For example, can one maintenance base reproduce a result (signal output and decision) that is the same as that produced at another base.

The REPEATABILITY of a procedure is generally characterized by process control and variances in application of the procedure, and includes “human factors” for those applications involving signal or pattern recognition by human operators. Repeatability is defined as the ability for a specific NDI technique to be used repeatedly on the same specimen and to obtain the same result.

Finally, accuracy and precision in DATA RECORDING are required to provide confidence in the data provided.

Probability of Detection (POD) methodology was initially developed to assess and validate inherent capabilities of various non-destructive inspection (NDI) procedures. Reproducibility and repeatability were assumed and output variances were attributed to “human factors”. Precision in crack size measurement and documentation was required to minimize variances in NDI output (capability) as a function for crack size. Rigor and confidence in the detection process required a significant number of detection opportunities (trials) to characterize and quantify the detection output. Detection was and is generally recorded as a “HIT OR MISS” (detect or failure to detect) output. The basis for detection (detection threshold) was assumed to be

ANNEX A – DATA COLLECTION PROCESS

constant. Good engineering practice and economics required that the detection threshold must result in a low level of “false calls” (a detection call when no crack is present).

Probability of Detection (POD) methodology requires passing a large number of cracks or other anomalies (typically 60 or more) through an NDI process and recording the results as “HIT OR MISS” or as a scalar quantity with respect to actual crack size. The resulting data is then analyzed and fit to a cumulative log-normal model, as is discussed in Section 5.2 of the main report. Figure A-1 shows a typical POD curve.

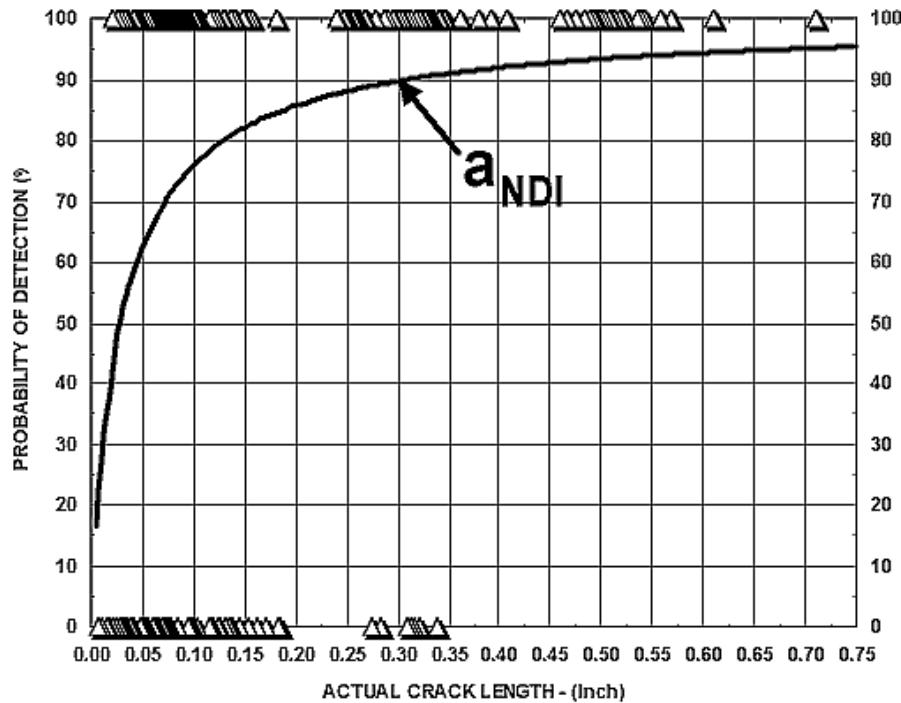


Figure A-1: A Typical Probability of Detection (POD) Curve.

Wide-spread use of the **POD** methodology to characterize, quantify and validate **NDI** procedure capabilities has identified significant variance in both **REPRODUCIBILITY** and **REPEATABILITY** due to variances in “calibration” and equipment/probe/transducer/inspection materials performance. It follows that the greater the variance in the **REPRODUCIBILITY** and **REPEATABILITY**, the greater the variance in applied NDI procedures and the resultant POD output. This has been one of the key obstacles to acceptance of the POD methodology – experiments for POD estimation must account for the expected variances at the level of implementation of the technique, not just at the laboratory level. Annex C provides an example of variances in reproducibility and repeatability in a practical maintenance situation.

In addition to challenges of variances in **REPRODUCIBILITY** and **REPEATABILITY** in applied **NDI** procedures, **POD** characterization from maintenance data involves additional challenges in precision, in sizing the detected anomalies at the time of the **NDI** procedure application and an absence of crack sizes for “missed anomalies”. The fidelity and usefulness of **POD** performance characterization from maintenance data is therefore dependent on variances in data quality (variance bounds). Variance in the quality of recorded data may result in variances in POD that neither reflect an accurate or useful capability of an **NDI** procedure.

For purposes of characterizing applied NDI maintenance procedures by the POD method, useful data must include attention to and consideration of:

- precision in crack size measurements,
- precision in “calibration”, and
- precision in process control in procedure application.

The quality of the data is characterized by precision in those three parameters in data collection and reporting/recording. Although some output can be gleaned from lesser quality data, the fidelity, applicability and usefulness of the POD output is reduced.

A.1.1 Precision in Crack Size Measurement / Actual Crack Size Measurements

The most useful information that can be added to NDI detection (HIT) data is that of physical measurement of actual crack size. Independent actual crack size measurement is a good practice to validate the NDI detection (and document FALSE CALLS), and to provide an important measure of NDI measurement process control. Precision in the independent measurement provides increased fidelity of the data for purposes of life cycle system management. Documentation is typically that of crack length or crack depth. An assumed crack aspect ratio is often used to estimate crack depth from documented surface crack length.

For those NDI methods involving visual inspection of part surfaces (such as visual, liquid penetrant and magnetic particle methods), direct surface crack length measurements may be made and documented. For those NDI methods involving an electronic output, comparison of the response from a crack in a test object to that from a “calibration artefact” is often the value recorded, and the quality of the measurement is dependent on both the fidelity of the recorded electronic output and on the quality and measured precision of the “calibration artefact”. The precision and accuracy of the measured/recorded output, in terms of “crack size”, is a primary factor in data quality and in data usefulness in POD quantification.

Surface crack length is ideally measured under load with optical magnification to a precision of ± 0.001 inch (0.0254 mm). For “calibration artefact” and laboratory test specimens, such measurement can be made rapidly and economically. For field applications, surface crack length may be measured under magnification, may be estimated by the use of an optical reticule in a hand-held magnifier, may be estimated by the judgment of the operator, or may be inferred from the step reamer used to remove the eddy current indication in a fastener hole. The greater the variance in the measurement, the lower the fidelity of any resulting **POD** analyses. It is estimated that a 3% error in POD may result from measurement tolerance of ± 0.005 inches.

Internal cracks are ideally characterized by breaking the cracks open and measuring actual crack size by metallographic methods. Such documentation is typically used for controlled characterizations using fabricated test specimens, but may also be provided on a sampling basis associated with process characterization and/or failure analysis. Such characterization may be used for the production of “calibration artefacts” by replicating samples, repeated measurements and documentation of all samples, fractures and measurements of one specimen in each replicated sample pair.

A more common method is to use side-drilled or flat-bottom holes for purposes of “calibration” and to relate responses to those from characterized cracks which are broken open and measured. Measurement precision to ± 0.001 inch (0.0254 mm) is easily provided by metallographic methods. Alternative measurement methods and precision may be used, but the method and precision must be recorded for later use in estimating errors and for use in data pooling.

ANNEX A – DATA COLLECTION PROCESS

In the absence of a quantified measurement of actual crack size, measurement and recording of signal and noise responses from individually detected anomalies and its relative response to a “calibration artefact” is useful information in both an indicator of data quality and a factor to be considered in “data pooling”. Use of the same type of “calibration artefact” is often considered to be sufficient to provide consistency in both detection and measurement. Unfortunately, variance in response between artefacts at various field locations is often unknown and variance in results is unknown. Such data are useful, but may result in wide variance in both POD and in consideration for data pooling.

The actual internal flaw size detection is often not known and judgment must be applied to both use of such data and in “pooling” such data from various sources. Fortunately, surface crack length is most often the basis for structural integrity assessments on airframes and engines.

Summary

Accuracy and precision in the measurement and recording of detected crack sizes will significantly affect the usefulness of the data in structural integrity assessments and the variance in threshold detection output as provided by POD analyses. Physical measurement of detected cracks is necessary to provide accuracy in POD analyses. This is an additional requirement in most maintenance NDI operations.

A.1.2 NDI Procedure Inherent Capability

The ultimate output of a **POD** assessment is to quantify applied **NDI** procedure crack detection capability. The inherent capability of an NDI procedure is characterized by a causal relationship between crack size and its relative signal response (output). A typical causal response relationship is shown schematically in Figure A-2. This model (and most NDI procedures) assume a monotonically increasing NDI response with increasing crack size. In order for the response relationship to be useful, the output must be capable of discriminating between responses from non-crack sources inherent in the detection/measurement application (signal/noise). Non-crack responses are typically termed application **NOISE** and may be due to test object surface roughness, grain size, impurities, stress state, etc. and should not be confused with “electronic noise” that is familiar in other applications (electronic noise is negligible when compared with other response sources).

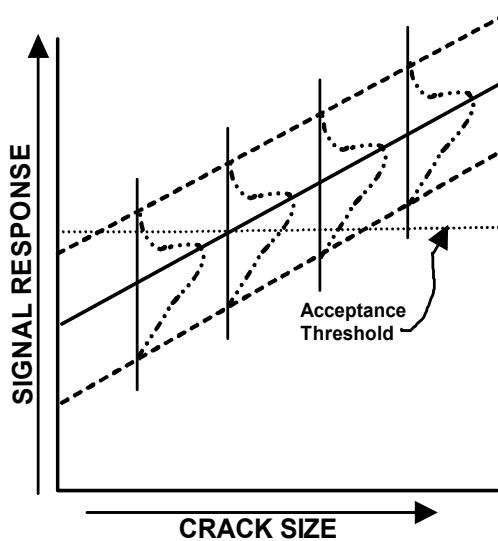


Figure A-2: Causal Relationship between NDI Signal Response and Crack Size.

A.1.3 Signal/Noise Response Relationships

When repetitive measurements of a single crack are made by an NDI procedure, a distribution of response values from the crack are generated that are similar to those produced in classical mechanical measurement methods. Simultaneously, a lower-level signal (background) response is generated that is characteristic of the surface condition, surface texture, grain structure, stress state, etc. of the test object. This background response is termed “NOISE”. A typical response from experimental measurements from a single crack is shown in Figure A-3.

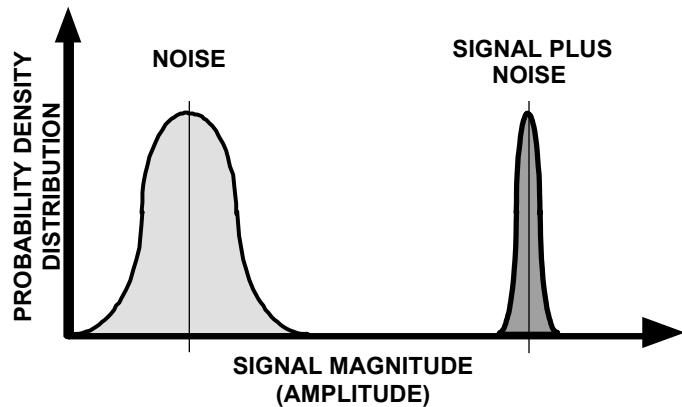


Figure A-3: Repetitive Responses from a Single Crack.

Repetitive response from multiple cracks of equal size results in broadening of the response distribution. This broadening is the result of crack-to-crack variations as well as measurement variations and are accounted for by using multiple cracks in the generation of a typical POD curve. The spread between the upper limit of the noise and the lower limit (signal and noise) of the crack response enables repetitive detection and discrimination/identification of cracks of that size without false calls (Type II errors). The practical threshold detection and discrimination limit is at that small crack size at which the signal and noise responses converge without overlap and detection/discrimination can be attained. It is wise to maintain a signal/noise margin (safety factor) in practical applications to allow for unanticipated variations in the NDI procedure.

Small sample sizes assume that the cracks selected are representative of the population of cracks to be detected and that crack-to-crack variance in application is bounded by the cracks selected for assessment. Various thrusts have been directed to modeling the crack-to-crack variance and have been successful for simple crack configurations. In many applications, this variance is accounted for by including a margin (safety factor) in detection requirements and by follow-up data collection and analysis of signal responses from service hardware. For complex configurations, larger margins may be used to address difficulties in validating margin assumptions.

When a small sample size is accepted as being representative of the population, repetitive measurements can be made on the selected cracks to establish the signal/measurement variance for each crack size sampled. Figure A-4 illustrates the broadening of response due to multiple measurements from cracks of equal size. This method produces a data set that can be used to establish a discrimination threshold and for plotting a POD curve. Although the number of measurements can provide a high measurement confidence level based on the small sample set, the measurements are not fully independent and thus less rigorous than that obtained by independent measurements on independent cracks. The POD curve generated is a measurement curve and

ANNEX A – DATA COLLECTION PROCESS

is representative of most important elements of the characterization task, but may not fully describe a capability if the response from service cracks of equal size varies significantly from the selected small sample set.

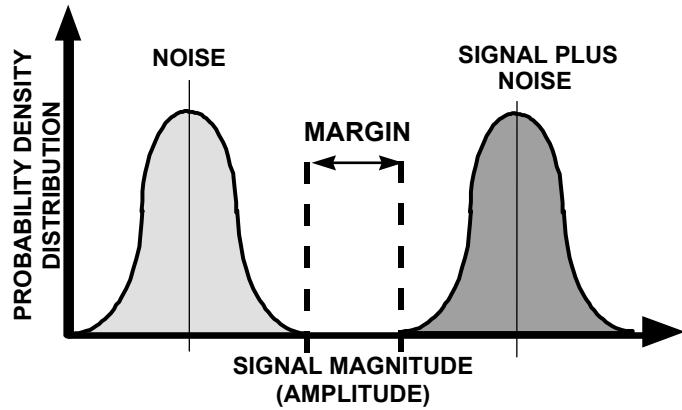


Figure A-4: Repetitive Response from Cracks of Equal Size.

The second part of NDI procedure optimization is in setting the acceptance level for the signals provided. For large cracks, the signal and noise are well separated and the threshold decision level can be easily set to provide clear discrimination as shown schematically in Figure A-5.

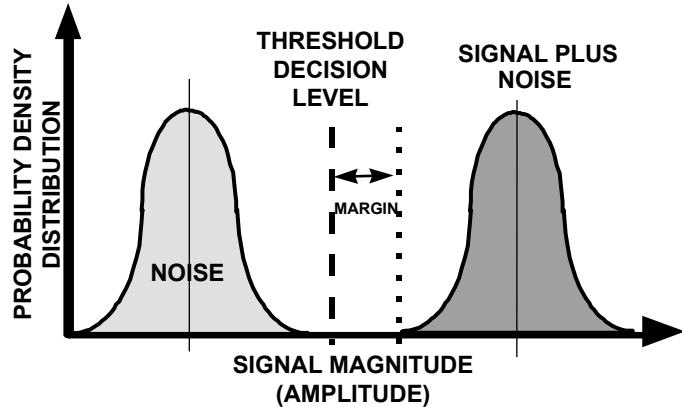


Figure A-5: Signal and Noise Separation for Large Cracks which Provide Clear Signal Discrimination.

If the threshold decision level is set too high, cracks will be missed. This condition may be imposed when signal and noise separation would otherwise allow clear discrimination as illustrated in Figure A-6. This is a condition often experienced when the threshold signal level is set on a slot in a “calibration” specimen and consideration is not given for the reduced response of a crack of a size that is equal to that of the slot.

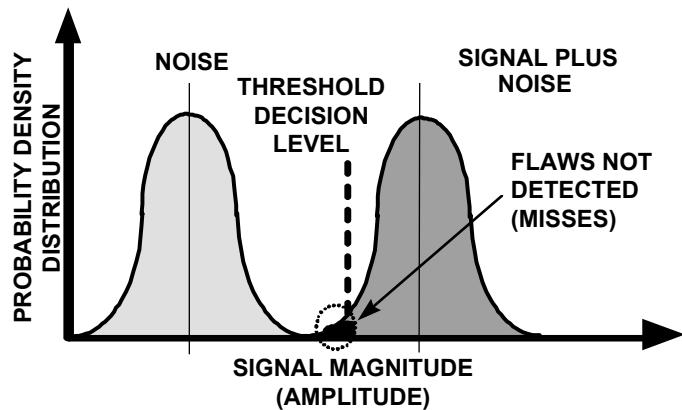


Figure A-6: High Threshold Level Results in Misses.

The limit of capability of an NDI procedure is reached when the signal and noise separation approaches zero. When signal and noise overlap, both misses and false calls will result as illustrated in Figure A-7. In this case, if the threshold level is set to assure detection, the number of false calls will increase. A level of false calls can be tolerated if a secondary procedure (usually NDI) is applied to resolve false calls and provide the required discrimination. **CAUTION:** Applying the same NDI procedure cannot resolve false calls since the same signal and noise conditions are equal. Likewise, an NDI procedure with lower discrimination capabilities does not provide resolution. This error has been frequently observed in the use of visual inspection to resolve penetrant findings.

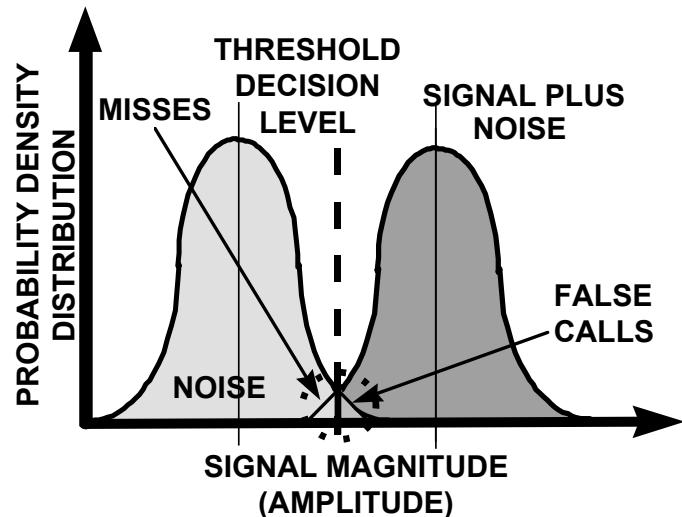


Figure A-7: Overlap of Signal and Noise Results in both Misses and False Calls.

The signal-to-noise response relationships define the practical ACCEPTANCE THRESHOLD for application of a specific NDI procedure. Typically a signal-to-noise ratio of 3-to-1 (response level from a crack of a threshold size to the response from the component in an area away from the crack – surface noise, grain noise, etc.) is required to produce discrimination at a practical level. The 3-to-1 signal-to-noise ratio takes into account crack-to-crack variances that are inherent to field applications. In Figure A-7, variance in signal

ANNEX A – DATA COLLECTION PROCESS

response at a given crack size is shown as a Gaussian distribution about a mean value. Increased precision in crack size measurement by the NDI procedure is accomplished by reduction in the signal variance at a crack size.

The causal model includes both detection of cracks with a response above the acceptance threshold (acceptance criteria) value and MISSES for responses below the acceptance threshold value. Maintenance NDI data provides the capability for documenting response data for detections (HITS), and thus, in itself, does not provide an adequate data set for purposes of generating a probability of detection (POD) curve. Estimation of the size of missed cracks may be made by back-calculation from the crack size detected at the next inspection interval using an assumed crack growth rate calculation method. In addition, maintenance data are often recorded only as detection (HIT); signal and noise response data are not provided and the detected crack size is assumed to be at the “calibration level”. Unfortunately, the threshold crack size detected is rarely at the “calibration level” and errors in the assumed crack size vary with the variance in signal response at the “calibration level” and with crack-to-crack response variance.

Summary

A useful causal response relationship between signal level and crack size is assumed to have been established during NDI procedure development and validation. In like manner, a constant acceptance threshold (detection) level is assumed to have been established during procedure validation and to have been further validated by field application experience. **CAUTION:** One consideration in ill-behaved data is the failure and or shift in the acceptance threshold or in the causal relationship.

A.1.4 NDI Procedure Reproducibility

REPRODUCIBILITY of a procedure is generally characterized by the inherent capability and variances in the procedure “calibration” process. If the instrument gain and response to given artefact can be reproduced, procedure reproducibility are demonstrated. It is assumed that the “foot print” of the probe/transducer, damping, frequency, gain corrections, etc. that are inherent to the procedure have been duplicated prior to “calibration” demonstration.

For NDE methods providing an electronic signal response, a single-point “calibration” is often used. Unfortunately, a single-point “calibration” is possible with NDE systems that provide significantly different responses. Figure A-8 illustrates variance in NDE response values for three cracks of different size (1, 2 and 3) resulting from an identical single-point “calibration” with NDE systems A, B and C having different response outputs. Such response variances are often due to differences in transducers, cabling and characteristic response of the instrument amplifier.

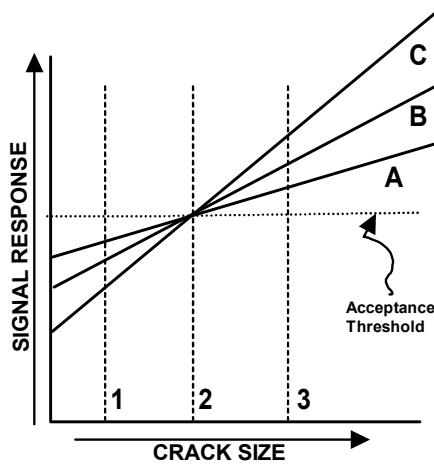


Figure A-8: Variations in Crack Response (1, 2 and 3) with a Variation in NDI System Response (A, B and C).

Although a single-point “calibration” may be adequate for NDE procedures applied to quality control, quantitative NDE requires reproducible response at three or more points that characterize the causal model for measurements in the required measurement range. The multiple-point “calibration” reduces variance in results at a single facility and may greatly reduce variance in results for measurements made at multiple facilities.

Edwards (see Annex C) has demonstrated variances in eddy current output with three and five-point reference measurements and has clearly demonstrated the need for reference to MASTER GAUGE artefacts, when an inspection is performed at multiple locations. The output response may not be linear with respect to artefact size as demonstrated in the rotating probe data, and “calibration” must include replication of the results produced for baseline validation of the NDI procedure.

Similar variances have been observed and are expected in ultrasonic and other measurement methods using electronic instruments.

A.1.4.1 Master Gauge “Calibration” Artefacts

It is difficult to provide “calibration” artefacts that provide an identical NDE response. All slots, notches or cracks of the same size do not provide the same NDE response. When the same NDE procedure is to be used at several locations or at a single location for an extended period of time, ‘master gauging’ is required to assure that all “calibration” artefacts provide the same response (or corrected response) as that of a “master gauge” that is preserved in a protected condition at a central location. The “master gauge” approach is highly recommended at a single facility due to potential damage or loss of the working “standard” artefact(s) and the resulting loss of traceability to the NDI procedure validation. This method is similar to that used in good metrology practice and is necessary to assure REPRODUCIBILITY of response at various locations and/or times.

The response to artefacts of equal size is compared to the response of the same size artefact in a “master gauge”. Correction factors are included with the working artefact to assure that the same response is obtained at all locations and inspection sequences. “Master gauge” artefacts of at least three different sizes are required to verify equal system response to service-induced cracks. Working “standards” should be periodically re-measured and responses re-verified in accordance with good quality assurance practices.

Summary

Data quality and data sets of differing quality are characterized by the rigor, care and objective assurance that NDI procedure “calibration” supports REPRODUCIBILITY in NDI detection and measurements to reduce/minimize this source of data variance. Lower data quality produced by variance in procedure REPRODUCIBILITY reduces both the POD capability for a procedure and the usefulness of the data in supporting structural integrity of the test object/system.

A.1.5 Repeatability and Process Control

REPEATABILITY in all NDE procedures is affected by rigid process control. Attention to and documentation of all elements of the NDE procedure and “calibration” procedure are required. Each NDI procedure should be documented in such detail that a second operator can set-up and repeat the procedure without questions. Typical NDI procedure documentation requirements are summarized below. In addition, both the REPRODUCIBILITY and REPEATABILITY / process control of an applied NDI procedure are dependent on HUMAN FACTORS. A short summary of HUMAN FACTOR effects on POD is summarized below. In the event that a change in a parameter is required, demonstration of equivalency to the previous procedure is required, including traceability to validation data.

For electronic NDE procedures, demonstration of equivalency may be by repetitive response measurements on reference cracks used in validation and made with repetitive “calibration” sequences.

For non-electronic NDE procedures such as fluorescent penetrant inspection, a full POD using validation cracks or a sub-set of the full POD set may be used to demonstrate equivalency of detection and discrimination. In addition, process control panels such as Testing and Monitoring (TAM) Panels may provide an indication of process control and procedure equivalency. Use of the same TAM panel for assessment of a “before” and “after” process change is required, since variations in TAM panels result in variations in out-put response. *Careful cleaning of the TAM panels between inspections sequences is required for such comparisons, as well as for daily use.*

A.1.6 NDI Procedure Documentation Requirements Summary

NDE aircraft maintenance data collection for purposes of quantifying NDE procedure capability, damage tolerance and residual life analysis requires the following items as a minimum (Table A-1).

Table A-1: Aircraft Maintenance NDI Data Collection Guidelines

Item	Description
1	Description of inspection area and characteristics associated with the inspection <ul style="list-style-type: none"> • Overall parameters • Critical parameters
2	Written NDE procedure including “calibration”
3	Reference data on validation of the written NDE procedure
4	Rigid process control in all procedures applications
5	Documented actual crack size to a precision of ± 0.001 inch (± 0.0254 mm) Documentation should include a record of “FALSE CALLS”

Reduced variance in NDE procedure application is strongly recommended. The following items (Table A-2) are recommended to reduce NDE measurement variance.

Table A-2: Recommended Procedures for the Reduction of NDI Measurement Variance

Item	Description
1	Document NDE signal response for each crack found
2	Rigorous validation of the NDE procedure including the cracks used and results
3	Three-point “calibration” for all electronic NDE procedures
4	Master gauge of all “calibration” artefacts

A.1.7 Human Factors Considerations

When a NDI procedure fails to detect a crack, the most frequent reason stated is HUMAN FACTORS/OPERATOR ERROR. Although attention must be given to operator training to transfer knowledge and to develop skill, the list of NDI procedure documentation requirements is daunting. Failure to detect defects may be due to:

- Flaw (Artefact) Variables
- Test Object Variables
- NDE Method Variables
- NDE Materials Variables
- NDE Equipment Variables
- NDE Procedure Variables
- NDE Process Variables including environment
- Calibration Variables
- Acceptance Criteria / Decision Variables
- Human Factors

Unless the preceding variables are under control, the operator at the end of the list has little chance of detection. Some of the variables are controlled by the operator, as is evident from the list of NDI procedure requirements. Other variables are beyond operator control. For example, facility variables are rarely recorded as a part of NDI data documentation.

The dominant operator dependent factor on POD capability is recency of experience with the specific test object and NDI procedure application. A trial run with known artefacts, to sharpen operator skills before an inspection is initiated, is much more beneficial than is additional classroom training or a written examination at a central facility.

A.1.8 Summary

The NDI procedure should be documented in sufficient detail to enable repetition of the measurements made in validation of the NDI procedure. Knowledgeable and skilled operators are essential to the measurement

process, and variances in the procedure or in operator skill will be reflected in variances in POD. In most cases, a judgment call must be made on the quality of data produced by application of an NDI procedure and on the relative skill of the operator applying the procedure. It is obvious that similarities in both data quality and operator skill must be considered as a factor in maintenance data “pooling”.

A.2 DATA DOCUMENTATION

Documentation of a data set must reflect consideration of the factors and parameters discussed herein. A judgment call must be made concerning the quality of the data and of back-calculations using flaw growth analysis to generate “Misses” for the data set. Application of a “handbook” procedure at different facilities does not assure that the data quality or capabilities of different facilities are equal. This is particularly applicable to data “pooling” for purposes of adding additional detection opportunities and trials. Pooling of data of differing quality degrades the quality of the combined data set (analogous to adding stones to the soup). Tables A-3 and A-4 document basic procedure and reporting data which should be recorded for any NDI procedure, and which are required in order to pool data with confidence.

Table A-3: Standardised NDI Procedure – Basics

1	Procedure no. (unique) and issue
2	Requirements to inspector level
3	Component to be examined
4	Area to be examined
5	Purpose of examination
6	Equipment required
7	Aircraft and part preparation
8	Calibration and sensitivity
9	Procedure
10	Acceptance criteria
11	Reporting
12	Man-hours of inspection
13	Additional information
14	Issuing organisation
15	Date of issue and pages included
16	Sign for approval
17	Detailed drawings of inspection area (including possible defects)

Table A-4: Standardized NDI Report Form

1	Issuing organisation
2	Report number (unique)
3	Date of inspection
4	National approval (if existing)
5	Type of inspection (ET bolt hole, ET surface, UT, etc.)
6	Type of aircraft
7	Serial no. or Tail no.
8	Flight hours (if necessary)
9.1	Inspected part
9.2	P/N
9.3	S/N
10	Related inspection procedure and actual issue
11	Other related documents
12	Inspected material (Alu, steel, CFRP, etc.)
13	Surface condition (blanc, painted)
14	Equipment used, Manufacturer (including probes, etc.)
15	Actual deviations to inspection procedure
16	Remarks to inspection conditions
17	Actual findings
18	Drawing of findings
19	Defect size, position and orientation; and method of sizing
20	Remarks to defects
21	Acceptable/not acceptable referring to inspection procedure
22	Place of inspection
23	Name, stamp and sign of inspector

Examples of the documentation required as a minimum for data pooling are shown in Table A-5. A more complete set of information required for the development and documentation of inspection procedures is provided in Tables A-6 and A-7, courtesy of Daimler-Chrysler Aerospace.

ANNEX A – DATA COLLECTION PROCESS

Table A-5: Typical Required NDI Procedure Documentation (from NDE Capabilities Databook – Third Edition)

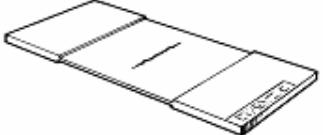
A 1000(1L)	DATA SET DESCRIPTION (ET - 01 (1) CRACK LENGTH)
METHOD:	Eddy Current
TEST OBJECT TYPE:	Flat Plate - 3.5 inches by 16 inches, cracks on both sides
NDE PROCEDURE:	Eddy Current - Contact Probe 100 kHz, Meter Readout
ARTIFACT TYPE:	Fatigue Cracks - $R < 0.70$ (Shaped EDM starter notch initiation, growth in bending and tension / tension
ARTIFACT SHAPE:	ASPECT RATIO - 0.1 TO 0.5 (a/2c) --- DEPTH TO THICKNESS - 0.2 TO 0.5 (a/t)
ARTIFACT VERIFICATION:	Destructive analysis and measurement
MATERIAL:	2219 Aluminum T-87
TEST OBJECT THICKNESS:	0.060 and 0.225 inch nominal
TEST OBJECT CONDITION:	-01, "As Machined", -02, "After Etch", -03, B1 "After Proof"
SURFACE FINISH:	125 and 32 RMS - representative of good machining practices
APPLICATION:	Hand Scanning - Manual Readout
DATA SET IDENTIFIER:	ETAAA01-A,B,C; ETAAA02-A,B,C; ETAAA03-A,B,C
TYPE OF DATA:	Hit / Miss with estimated crack lengths
TEST OPPORTUNITIES:	311 Cracks
DETECTED:	ETAAA01-A= 208, B= 224, C= 205; 02-A= 228, B= 273, C= 243; 03-A= 264, B= 268, C= 266
FALSE CALLS:	Not reported
REFERENCE:	NASA CR-2369 Rummel, Ward D., Paul H. Todd Jr., Sandor A. Frecka, and Richard A. Rathke, The Detection of Fatigue Cracks by Nondestructive Testing Methods , February 1974.
DATE:	November 1971 - June 1973
WORK SPONSOR:	W.L. Castner, NASA Lyndon B. Johnson Space Center
PERFORMING ORGANIZATION:	Martin Marietta Aerospace, Denver, Colorado
NOTES:	<p>This program was performed in support of the National Aeronautics Administration (NASA) Space Shuttle design and is the first known publication of nondestructive evaluation data in a continuous function probability of detection (POD).</p> <p>Flaws were induced in 105 panels (both sides). Thirteen blank panels were included for a total of 118 panels.</p> <p>The original data analysis was in the form of a moving average plot. Data have been reanalyzed and plotted here by the maximum likelihood / log logistic method.</p> <p>A parallel program was conducted by the General Dynamics Corp, San Diego, CA.; test panels were exchanged and inspections repeated by both organizations.</p>
	90% POD Length - "AS MACHINED" "AFTER ETCH" "AFTER PROOF" A= 0.196 in. A= 0.198 in. A= 0.052 in. B= 0.184 in. B= 0.071 in. B= 0.037 in. C= 0.295 in. C= 0.270 in. C= 0.0871 in.

Table A-6: Overall Parameters to Define a Characteristic Inspection

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Cause for Inspection:</u> – Inspection history / Defect history – Type of defect – Risk on-going by the defect	<u>Cause for Inspection:</u> – Inspection history / Defect history – Type of defect – Risk on-going by the defect	<u>Cause for Inspection:</u> – Inspection history / Defect history – Type of defect – Risk on-going by the defect	<u>Cause for Inspection:</u> – Inspection history / Defect history – Type of defect – Risk on-going by the defect
<u>Affected Aircraft, Component, Part, P/N, Material</u> – Aircraft modifications present? – Variation in material, geometry, access, sensitivity	<u>Affected Aircraft, Component, Part, P/N, Material</u> – Aircraft modifications present? – Variation in material, geometry, access, sensitivity	<u>Affected Aircraft, Component, Part, P/N, Material</u> – Aircraft modifications present? – Variation in material, geometry, access, sensitivity	<u>Affected Aircraft, Component, Part, P/N, Material</u> – Aircraft modifications present? – Variation in material, geometry, access, sensitivity
<u>Time of Inspection</u> – After hard landing – Periodically – Maintenance level – Applicability of alternative inspection if primary inspection not possible	<u>Time of Inspection</u> – After hard landing – Periodically – Maintenance level – Applicability of alternative inspection if primary inspection not possible	<u>Time of Inspection</u> – After hard landing – Periodically – Maintenance level – Applicability of alternative inspection if primary inspection not possible	<u>Time of Inspection</u> – After hard landing – Periodically – Maintenance level – Applicability of alternative inspection if primary inspection not possible
Required NDI-personal qualification	Required NDI-personal qualification	Required NDI-personal qualification	Required NDI-personal qualification
<u>Necessary NDI Equipment</u> – Type of equipment – Type of surface probe (shielded, 90°, flexible shaft, diameter) – Special tooling (probe guides, spring loads, printer, handling aids)	<u>Necessary NDI Equipment</u> – Type of equipment – Type of rotating probe (spreading heat, length, diameter) – Special tooling (probe guides, spring loads, printer, handling aids)	<u>Necessary NDI Equipment</u> – Type of equipment – Type of UT-probe (diameter, MHz, delay line, adapted delay lines, focussed, couplant) – Special tooling (probe guides, spring loads, printer, handling aids)	<u>Necessary NDI Equipment</u> – Type of equipment – Type of UT-probe (diameter, MHz, wedge angle, special form of delay line, location of connector, outer size) – Special tooling (probe guides, spring loads, printer, handling aids)

ANNEX A – DATA COLLECTION PROCESS

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Procedure</u> <ul style="list-style-type: none"> – Actual revision of procedure – Language – Other procedures affected or referred – Misc. (pages, issue, issuer) – Applicability 	<u>Procedure</u> <ul style="list-style-type: none"> – Actual revision of procedure – Language – Other procedures affected or referred – Misc. (pages, issue, issuer) – Applicability 	<u>Procedure</u> <ul style="list-style-type: none"> – Actual revision of procedure – Language – Other procedures affected or referred – Misc. (pages, issue, issuer) – Applicability 	<u>Procedure</u> <ul style="list-style-type: none"> – Actual revision of procedure – Language – Other procedures affected or referred – Misc. (pages, issue, issuer) – Applicability
<u>Preparation</u> <ul style="list-style-type: none"> – Access/removed parts – Surface (blanc, paint removal, blistered paint) – Paint thickness measurement 	<u>Preparation</u> <ul style="list-style-type: none"> – Access/removed parts – Surface (blanc, paint removal, blistered paint) – Fastener removal 	<u>Preparation</u> <ul style="list-style-type: none"> – Access/removed parts – Surface (blanc, paint removal, blistered paint) – Paint thickness measurement 	<u>Preparation</u> <ul style="list-style-type: none"> – Access/removed parts – Surface (blanc, paint removal, blistered paint) – Paint thickness measurement
<u>Calibration Standard</u> <ul style="list-style-type: none"> – Shape – Material – Thickness – Surface treatment – Coatings – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer) – Layers – Spacings – Identification 	<u>Calibration Standard</u> <ul style="list-style-type: none"> – Shape – Material – Thickness – Surface treatment – Coatings – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer) – Layers – Spacings – Identification 	<u>Calibration Standard</u> <ul style="list-style-type: none"> – Shape – Material – Thickness – Surface treatment – Coatings – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer) – Layers – Spacings – Identification 	<u>Calibration Standard</u> <ul style="list-style-type: none"> – Shape – Material – Thickness – Surface treatment – Coatings – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer) – Layers – Spacings – Identification
<u>Calibration</u> <ul style="list-style-type: none"> – Equipment set-up – Probe connection – Warm-up time – Basic settings (gain, MHz, x-y/y-x-display, time deflection, x-/y-gain, filter) – Signal orientation 	<u>Calibration</u> <ul style="list-style-type: none"> – Equipment set-up – Probe connection – Warm-up time – Basic settings (gain, MHz, x-y/y-x-display, time deflection, x-/y-gain, filter) – Signal orientation 	<u>Calibration</u> <ul style="list-style-type: none"> – Equipment set-up – Probe connection – Warm-up time – Basic settings (gain, filter, delay, range zoom, DAC) – Defect of calibration standard to be used 	<u>Calibration</u> <ul style="list-style-type: none"> – Equipment set-up – Probe connection – Warm-up time – Basic settings (gain, filter, delay, range zoom, DAC) – Defect of calibration standard to be used

ANNEX A – DATA COLLECTION PROCESS

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Calibration</u> (continued from previous page) <ul style="list-style-type: none"> – Defect of calibration standard to be used 	<u>Calibration</u> (continued from previous page) <ul style="list-style-type: none"> – Defect of calibration standard to be used 	<u>Calibration</u> (continued from previous page) <ul style="list-style-type: none"> – Threshold (start/end, shape, light, acoustic, trigger) – Signal dynamic – Report on calibration – Repetition after parts of inspection 	<u>Calibration</u> (continued from previous page) <ul style="list-style-type: none"> – Threshold (start/end, shape, light, acoustic, trigger) – Signal dynamic – Report on calibration – Repetition after parts of inspection
<u>Localisation and Definition of Inspection Area</u> <ul style="list-style-type: none"> – Drawing of aircraft – Drawing of component – Drawing of inspection area – Drawing of scans – Remarks on geometry, material, restrictions, precautions, other influences 	<u>Localisation and Definition of Inspection Area</u> <ul style="list-style-type: none"> – Drawing of aircraft – Drawing of component – Drawing of inspection area – Drawing of affected bolt holes – Remarks on geometry, material, restrictions, precautions, other influences – Affected layer 	<u>Localisation and Definition of Inspection Area</u> <ul style="list-style-type: none"> – Drawing of aircraft – Drawing of component – Drawing of inspection area – Drawing of scans – Remarks on geometry, material, restrictions, precautions, other influences – Affected layer 	<u>Localisation and Definition of Inspection Area</u> <ul style="list-style-type: none"> – Drawing of aircraft – Drawing of component – Drawing of inspection area – Drawing of scans – Remarks on geometry, material, restrictions, precautions, other influences – Affected layer
<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Identification of parts of the inspection areas – Visual inspection of surface treatment/status/condition – Marking of inspection areas – Material identification – Scanning of flat/curved areas – Testing of edges, radii, gaps, holes, radii (inner/outer) – Consider changing in materials, plating, paint, thickness, ferro-magnetic changing 	<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Identification of parts of the inspection areas – Visual inspection of surface treatment/status/condition – Marking of inspection areas – Material identification – Record signals – Use proper handling tools – Cracked layer – Crack depth – Crack orientation – Crack length / crack start 	<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Identification of parts of the inspection areas – Visual inspection of surface treatment/status/condition – Marking of inspection areas – Material identification – Record signals – Use proper handling tools – Cracked layer – Crack depth – Crack orientation – Crack length / crack start 	<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Identification of parts of the inspection areas – Visual inspection of surface treatment/status/condition – Marking of inspection areas – Material identification – Record signals – Use proper handling tools – Cracked layer – Crack depth – Crack orientation – Crack length / crack start

ANNEX A – DATA COLLECTION PROCESS

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Inspection</u> (continued from previous page) <ul style="list-style-type: none"> – Awareness of tilting, spacing, guiding, etc. the probe – Record signals – Use proper handling tools – Cracked layer – Crack depth – Crack orientation – Crack length – Signal interpretation – Consider signal dynamics, Z-positions, scanning matrix, scanning direction – Crack start – Crack amplitude related to threshold – Use backup-NDI when defects are found 	<u>Inspection</u> (continued from previous page) <ul style="list-style-type: none"> – Signed interpretation – Use proper probe diameter – Consider changing in diameter, material, gaps, spacers, nut retainer, hole length, layer thickness, corrosion, tapered shape, depths and grooves in the hole, etc. – Record phase shift – Crack amplitude related to threshold – Use back-up NDI when defects are found 	<u>Inspection</u> (continued from previous page) <ul style="list-style-type: none"> – Signed interpretation – Crack amplitude related to threshold waves, reduced resolution in different depth, etc. – Use proper delay line – Take care about couplant – Record equivalent artificial defect – Consider changing in thickness material, paint thickness, inner and outer geometry, beam scattering, beam reflection, beam deflection, absorption, additional waves, splitter waves, reduced resolution in different depth, etc. – Use back-up NDI when defects are found 	<u>Inspection</u> (continued from previous page) <ul style="list-style-type: none"> – Signed interpretation – Crack amplitude related to threshold – Consider changing in thickness material, paint thickness, inner and outer geometry, beam scattering, beam reflection, beam deflection, absorption, additional waves, splitter waves, reduced resolution in different depth, etc. – Use proper delay line – Take care about couplant – Record equivalent artificial defect (i.e. flat-bottom hole) – Use back-up NDI when defects are found
<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls 	<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls 	<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls 	<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls
<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Make decision for further use/ removal/repair of the part – Mark defect durable on part – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.) 	<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Make decision for further use/ removal/repair of the part – Mark defect durable on part – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.) 	<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Make decision for further use/ removal/repair of the part – Mark defect durable on part – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.) 	<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Make decision for further use/ removal/repair of the part – Mark defect durable on part – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.)

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Documentation</u> (continued from previous page) <ul style="list-style-type: none"> – Report directly to ground staff, material office, stress department, etc. – Store records until 	<u>Documentation</u> (continued from previous page) <ul style="list-style-type: none"> – Report directly to ground staff, material office, stress department, etc. – Store records until 	<u>Documentation</u> (continued from previous page) <ul style="list-style-type: none"> – Report directly to ground staff, material office, stress department, etc. – Store records until 	<u>Documentation</u> (continued from previous page) <ul style="list-style-type: none"> – Report directly to ground staff, material office, stress department, etc. – Store records until
<u>Reassembly/Final Treatment</u> <ul style="list-style-type: none"> – Treat part by painting, coatings, etc. – Re-install fasteners, other parts, etc. 	<u>Reassembly/Final Treatment</u> <ul style="list-style-type: none"> – Treat part by painting, coatings, etc. – Re-install fasteners, other parts, etc. 	<u>Reassembly/Final Treatment</u> <ul style="list-style-type: none"> – Treat part by painting, coatings, etc. – Re-install fasteners, other parts, etc. 	<u>Reassembly/Final Treatment</u> <ul style="list-style-type: none"> – Treat part by painting, coatings, etc. – Re-install fasteners, other parts, etc.
<u>Additional Information</u> <ul style="list-style-type: none"> – Will cracks not be reworked? – Will same NDI-people repeatedly do this NDI on the same part and defect? – Are problems rising during inspection? – Is there special pressure on NDI-specialist? – Is the equipment and standard the “same” or equivalent? – Are there disturbing effects (i.e. dirt not removed, hot in hangar, cold outside, work inside, fuel tank, etc.)? – How often does the NDI-specialist do this type of inspection or similar ones? – How often do they inspect at all? 	<u>Additional Information</u> <ul style="list-style-type: none"> – Will cracks not be reworked? – Will same NDI-people repeatedly do this NDI on the same part and defect? – Are problems rising during inspection? – Is there special pressure on NDI-specialist? – Is the equipment and standard the “same” or equivalent? – Are there disturbing effects (i.e. dirt not removed, hot in hangar, cold outside, work inside, fuel tank, etc.)? – How often does the NDI-specialist do this type of inspection or similar ones? – How often do they inspect at all? 	<u>Additional Information</u> <ul style="list-style-type: none"> – Will cracks not be reworked? – Will same NDI-people repeatedly do this NDI on the same part and defect? – Are problems rising during inspection? – Is there special pressure on NDI-specialist? – Is the equipment and standard the “same” or equivalent? – Are there disturbing effects (i.e. dirt not removed, hot in hangar, cold outside, work inside, fuel tank, etc.)? – How often does the NDI-specialist do this type of inspection or similar ones? – How often do they inspect at all? 	<u>Additional Information</u> <ul style="list-style-type: none"> – Will cracks not be reworked? – Will same NDI-people repeatedly do this NDI on the same part and defect? – Are problems rising during inspection? – Is there special pressure on NDI-specialist? – Is the equipment and standard the “same” or equivalent? – Are there disturbing effects (i.e. dirt not removed, hot in hangar, cold outside, work inside, fuel tank, etc.)? – How often does the NDI-specialist do this type of inspection or similar ones? – How often do they inspect at all?

ANNEX A – DATA COLLECTION PROCESS

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Additional Information</u> (continued from previous page)			
<ul style="list-style-type: none"> – Is it a routine job for them? – Are there physical limitations (i.e. access to do this inspection)? – Is the proposed time enough for a thorough inspection? – Do they know the limits of this inspection technique? – Do they know the consequences when a dramatic failure occurs? – Is there a motivation to find a crack or does this cause additional work and stress to them? – Are there evaluation limits available, where no action is necessary? 	<ul style="list-style-type: none"> – Is it a routine job for them? – Are there physical limitations (i.e. access to do this inspection)? – Is the proposed time enough for a thorough inspection? – Do they know the limits of this inspection technique? – Do they know the consequences when a dramatic failure occurs? – Is there a motivation to find a crack or does this cause additional work and stress to them? – Are there evaluation limits available, where no action is necessary? 	<ul style="list-style-type: none"> – Is it a routine job for them? – Are there physical limitations (i.e. access to do this inspection)? – Is the proposed time enough for a thorough inspection? – Do they know the limits of this inspection technique? – Do they know the consequences when a dramatic failure occurs? – Is there a motivation to find a crack or does this cause additional work and stress to them? – Are there evaluation limits available, where no action is necessary? 	<ul style="list-style-type: none"> – Is it a routine job for them? – Are there physical limitations (i.e. access to do this inspection)? – Is the proposed time enough for a thorough inspection? – Do they know the limits of this inspection technique? – Do they know the consequences when a dramatic failure occurs? – Is there a motivation to find a crack or does this cause additional work and stress to them? – Are there evaluation limits available, where no action is necessary?

Table A-7: Critical Parameters to Define a Characteristic Inspection

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Cause for Inspection:</u> – Type of defect			
<u>Affected Component, P/N, Area</u> – Aircraft modifications present?			
Required NDI-Personal Qualification	Required NDI-Personal Qualification	Required NDI-Personal Qualification	Required NDI-Personal Qualification

ANNEX A – DATA COLLECTION PROCESS

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Necessary NDI Equipment</u> – Type of surface probe (shielded, 90°, flexible shaft, diameter)	<u>Necessary NDI Equipment</u> – Type of rotating probe (spreaded heat, length, diameter)	<u>Necessary NDI Equipment</u> – Type of UT-probe (diameter, MHz, delayline, adapted delaylines, focussed, couplant)	<u>Necessary NDI Equipment</u> – Type of UT-probe (diameter, MHz, wedge angle, special form of delayline, location of connector, outer size)
<u>Procedure</u> – Actual revision of procedure	<u>Procedure</u> – Actual revision of procedure	<u>Procedure</u> – Actual revision of procedure	<u>Procedure</u> – Actual revision of procedure
<u>Preparation</u> – Access/removed parts – Surface (blanc, paint removal, blistered paint) – Paint thickness measurement	<u>Preparation</u> – Access/removed parts – Surface (blanc, paint removal, blistered paint) – Fastener removal	<u>Preparation</u> – Access/removed parts – Surface (blanc, paint removal, blistered paint)	<u>Preparation</u> – Access/removed parts – Surface (blanc, paint removal, blistered paint)
<u>Calibration Standard</u> – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer)	<u>Calibration Standard</u> – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer)	<u>Calibration Standard</u> – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer)	<u>Calibration Standard</u> – Defects included (manufacturing, type, size, length, shape, orientation, depth, layer)
<u>Calibration</u> – Basic settings (gain, MHz, x-y/y-x-display, time deflection, x-/y-gain, filter) – Signal orientation – Defect of calibration standard to be used – Threshold (level, start/end)	<u>Calibration</u> – Basic settings (gain, MHz, x-y/y-x-display, time deflection, x-/y-gain, filter) – Signal orientation – Defect of calibration standard to be used – Threshold (level, start/end)	<u>Calibration</u> – Basic settings (gain, filter, delay, range zoom, DAC) – Defect of calibration standard to be used – threshold (level, start/end) – Signal dynamic	<u>Calibration</u> – Basic settings (gain, filter, delay, range zoom, DAC) – Defect of calibration standard to be used – threshold (level, start/end) – Signal dynamic
<u>Localisation and Definition of Inspection Area</u> – Drawing of component – Drawing of inspection area – Drawing of scans	<u>Localisation and Definition of Inspection Area</u> – Drawing of component – Drawing of inspection area – Drawing of affected bolt holes – Affected layer	<u>Localisation and Definition of Inspection Area</u> – Drawing of component – Drawing of inspection area – Drawing of scans – Affected layer	<u>Localisation and Definition of Inspection Area</u> – Drawing of component – Drawing of inspection area – Drawing of scans – Affected layer

ANNEX A – DATA COLLECTION PROCESS

High Frequency Eddy Current Surface Inspection	Eddy Current Bolt Hole Inspection	Ultrasonic Longitudinal Wave Inspection	Ultrasonic Shear Wave Inspection
<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls 	<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls 	<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls 	<u>Special Remarks</u> <ul style="list-style-type: none"> – Discontinuities may cause false calls
<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Scanning of flat/curved areas – Testing of edges, bendings, gaps, holes, radii (inner/outer) – Consider changing in materials, platings, paint, thickness, ferromagnetic changings – Cracked layer – Crack depth / crack orientation – Crack length / crack start – Crack amplitude related to threshold 	<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Cracked layer – Crack depth – Crack orientation – Crack length / crack start – Use proper probe diameter – Crack amplitude related to threshold 	<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Cracked layer – Crack depth – Crack orientation – Crack length / crack start – Crack amplitude related to threshold 	<u>Inspection</u> <ul style="list-style-type: none"> – Deviations in sensitivity, threshold, etc. compared to calibration – Cracked layer – Crack depth – Crack orientation – Crack length / crack start – Crack amplitude related to threshold
<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.) 	<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.) 	<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.) 	<u>Documentation</u> <ul style="list-style-type: none"> – Attach records, drawings, information about the crack – Fill out attached/needed records (defect report, inspection report, datasheet, aircraft documentation, work order, etc.)
<u>Additional Information</u> <ul style="list-style-type: none"> – Are problems rising during inspection? 	<u>Additional Information</u> <ul style="list-style-type: none"> – Are problems rising during inspection? 	<u>Additional Information</u> <ul style="list-style-type: none"> – Are problems rising during inspection? 	<u>Additional Information</u> <ul style="list-style-type: none"> – Are problems rising during inspection?

Annex B – IMPROVED STATISTICAL ANALYSIS FOR SMALL DATA SETS

It is accepted that any POD data obtained from in-service inspection results will consist of a very limited number of data points, compared to data obtained via a dedicated POD trial. In order to make use of this data, it is important to have statistical analysis methods suitable for analysis of small data sets. Harding and Hugo (2003) present an alternative to the generally accepted analysis methodology given in Petrin, Annis and Vukelich (1993) – also found in MIL-HDBK-1823. Harding and Hugo (2003) use the same maximum likelihood estimation method, but employ an alternative chi-squared statistic to establish the 95% confidence limit curve. As discussed in Section 2.2 of the main report, inspection intervals are often based on a defect size with 90% probability of detection demonstrated with 95% statistical confidence ($a_{90/95}$). Thus, it is critical that valid methods are available for finding the 95% confidence limit using small data sets.

Section 5.2 of the main report outlined the POD(a) model (curve-fitting) approach described in USAF MIL-HDBK-1823. Maximum likelihood estimation is used to find parameter estimates $(\hat{\mu}, \hat{\sigma})$, which give the best fit to the observed data. The confidence limit curve is found by defining a confidence region, \mathcal{R} , in (μ, σ) space which is expected to contain the true values of the parameters μ, σ with a given confidence, Figure B-1(a). As the parameter vector θ varies within \mathcal{R} , the POD curves defined by $\text{POD}(a, \theta)$ will sweep out a band in the POD vs. a plane, Figure B-1(b). Thus the region \mathcal{R} defines a confidence band within which the entire true POD curve will lie with a given confidence level.

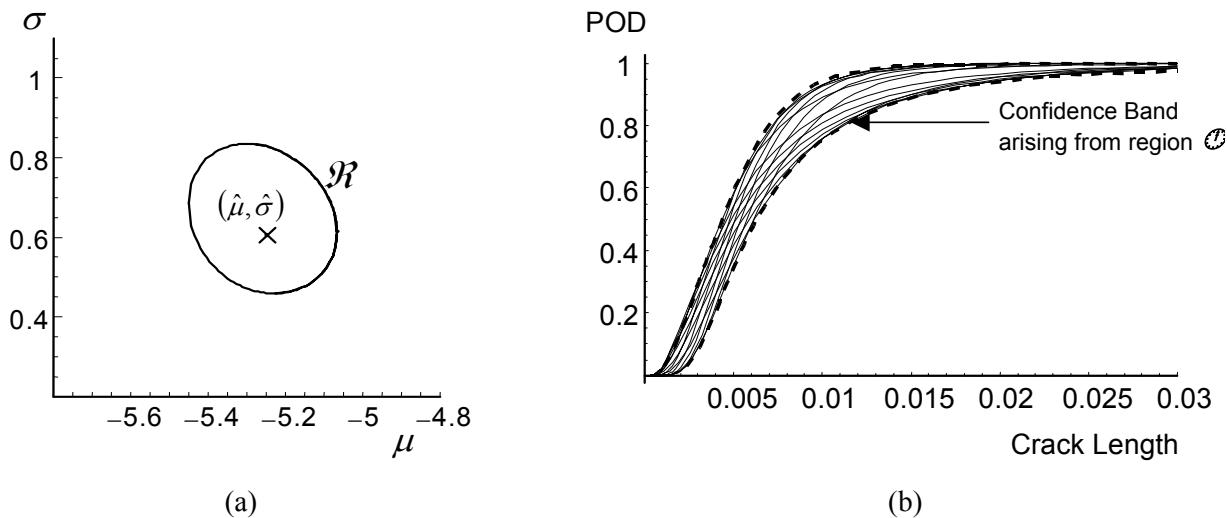


Figure B-1: (a) Confidence Region \mathcal{R} in (μ, σ) Space; (b) Confidence Band Defined by Confidence Region \mathcal{R} Contains all Possible POD Curves for (μ, σ) within \mathcal{R} .

The confidence region, \mathcal{R} , is defined by a statistic that is asymptotically chi-squared as the number of data points goes to infinity. Harding and Hugo have chosen a statistic that is better behaved for small data sets; for large data sets the two methods converge to the same result. Petrin, Annis and Vukelich use the following statistic, Q_1 , to define the confidence region for the parameter vector $\theta = (\mu, \sigma)$,

$$Q_1(\theta) = (\hat{\theta} - \theta)^T \mathbf{I}'(\hat{\theta})(\hat{\theta} - \theta) \quad (\text{B-1})$$

where $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$, $I'(\hat{\theta}) = -\partial^2 \ln L / \partial \theta_i \partial \theta_j$ and L is the likelihood function.

Harding and Hugo use the following statistic, Q_2 :

$$Q_2(\theta) = -2 \ln \left[\frac{L(\theta)}{L(\hat{\theta})} \right]. \quad (\text{B-2})$$

The boundary of the confidence region in each case is given by $Q(\theta) - \gamma = 0$, where γ is the critical chi-squared statistic.

The qualitative difference between methods Q_1 (Petrin, Annis and Vukelich) and Q_2 (Harding and Hugo) becomes evident when applied to small data sets, Figure B-2(a). The confidence region \mathcal{R}_2 defined by Q_2 follows a contour of the likelihood function and is frequently not centred on the parameter estimates $(\hat{\mu}, \hat{\sigma})$. By comparison, the form of Q_1 constrains the corresponding region \mathcal{R}_1 to be an ellipse centred on $(\hat{\mu}, \hat{\sigma})$.

The effect of the different shapes of \mathcal{R} on the lower confidence limit curves is shown in Figure B-2(b). For small data sets, \mathcal{R}_2 tends to be elongated in the direction of small μ and large σ compared to \mathcal{R}_1 . This part of the boundary corresponds to the lower confidence limit for high values of POD. The elongation of \mathcal{R}_2 in this direction gives a lower confidence limit which is significantly lower (more conservative) for Q_2 than Q_1 in the upper part of the curve above 50% POD. In the lower part of the curve (below 50% POD), the lower confidence limit given by Q_1 is more conservative. Q_2 exhibits the very useful behaviour that \mathcal{R}_2 becomes extremely elongated in the direction of large σ and small μ for data sets that contain too few hits or too many misses at large crack sizes to justify high values of POD with 95% confidence at any crack size. The corresponding lower confidence limit for Q_2 becomes horizontal at large crack sizes with a limiting maximum POD less than one. By comparison, the lower confidence limit given by Q_1 approaches a POD of one eventually at sufficiently large crack sizes, whatever the quality of the data set.

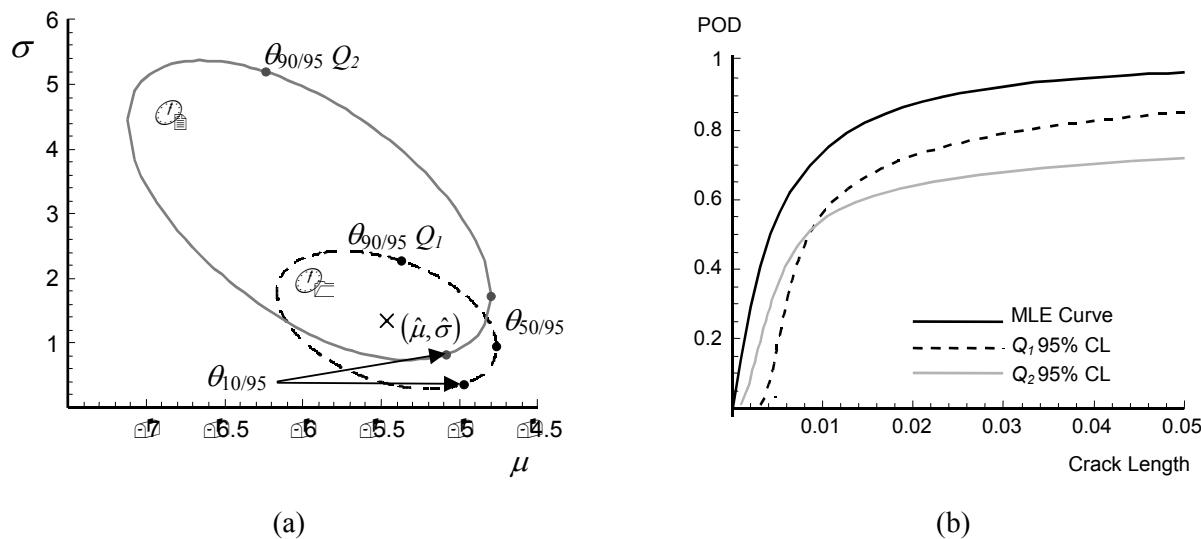


Figure B-2: (a) Boundary on Confidence Regions \mathcal{R}_1 and \mathcal{R}_2 Defined using Q_1 and Q_2 , respectively, and (b) Corresponding 95% Confidence Limit Curves, computed for same data set containing 50 ‘hit/miss’ inspections.

The behaviour of the confidence limits defined by Q_1 and Q_2 was explored for decreasing sample sizes using simulations comprising 2000 trials at each sample size. This large total number of trials was required to obtain a statistically significant number of non-conservative results. Figure B-3 plots against sample size the percentage of trials giving lower confidence limits which were non-conservative at any point on the curve. For Q_1 , the lower confidence limit curves become increasingly non-conservative as the sample size decreases below 200 data points. The lower confidence limit curves defined by Q_2 consistently maintain the expected non-conservative rate of 2.5%¹ down to data sets as small as 50 “hit/miss” observations.

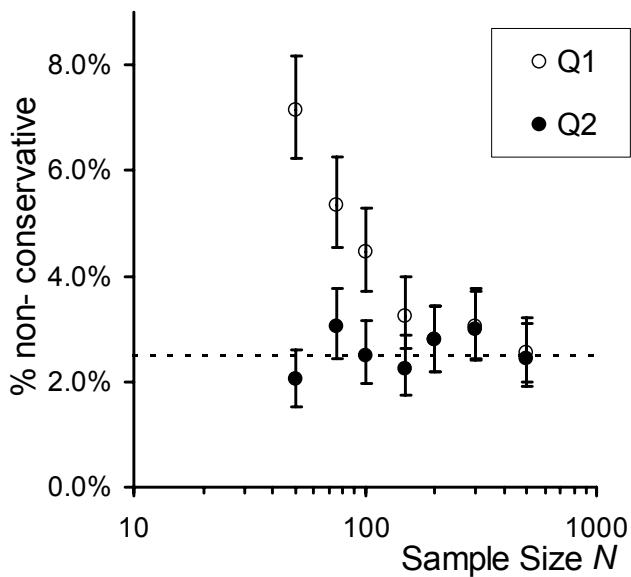


Figure B-3: Percentage of Trials with Q_1 and Q_2 Lower Confidence Limit Curves Non-Conservative at Any Point on the Curve, plotted as a function of sample size. Error bars denote the statistical uncertainty in the non-conservative rate based on a total of 2000 trials at each sample size. The dashed line at 2.5% denotes the expected percentage of non-conservative results.

The non-conservative rates for individual points on the lower confidence limit curves: $a_{10/95}$, $a_{50/95}$ and $a_{90/95}$ are examined in Figure B-4. Note that an individual point on the curve is expected to give a non-conservative rate significantly less than 2.5%. When using Q_1 , the values of $a_{10/95}$ are consistently more conservative than $a_{90/95}$ and this difference becomes more significant for smaller sample sizes. The high rate of non-conservative results for $a_{90/95}$ using Q_1 are of concern because $a_{90/95}$ is often the parameter of interest for setting safe inspection intervals. For Q_2 , the differences between non-conservative rates of $a_{10/95}$, $a_{50/95}$ and $a_{90/95}$ are much smaller and non-conservative rates below 2% are maintained for samples sizes down to 50 “hit/miss” observations.

¹ Note that the Q_1 and Q_2 methods give two-sided confidence limits with 95% confidence that no point on the true $POD(a)$ curve lies outside the band given by the upper and lower confidence curves. Consequently, the *lower* confidence limit is expected to be non-conservative with respect to the true POD at some point on the curve for 2.5% of trials at most.

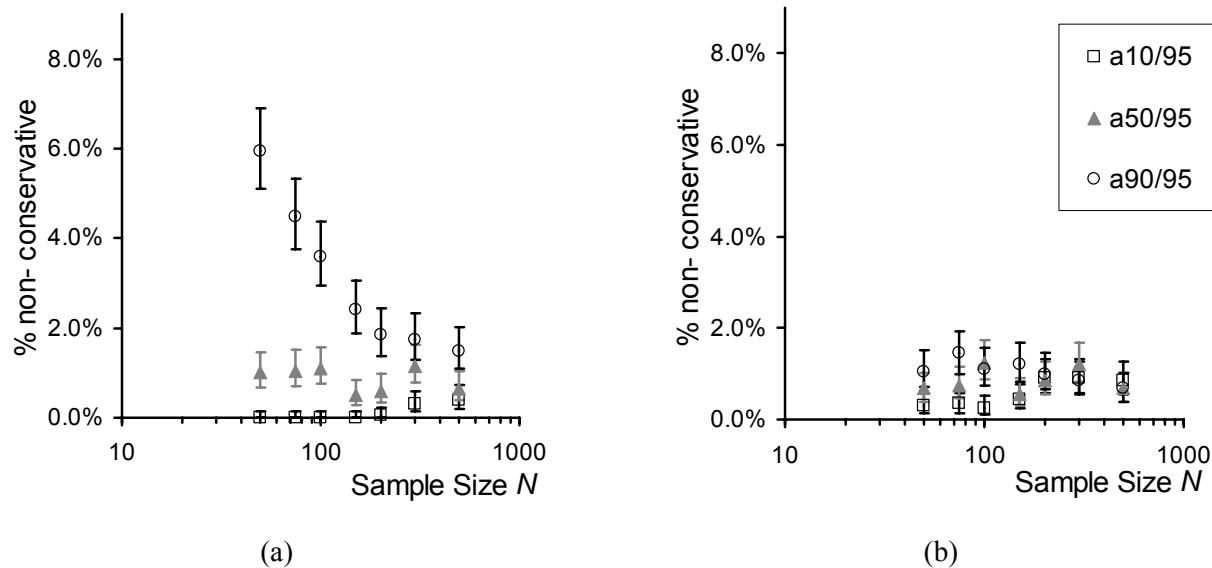


Figure B-4: Percentage of Trials with Q_1 and Q_2 Lower Confidence Limit Curves Non-Conservative at 10%, 50% and 90% POD, plotted as a function of sample size: (a) Q_1 method (b) Q_2 method. Error bars denote the statistical uncertainty in the non-conservative rate based on the total of 2000 trials at each sample size.

These results demonstrate that Q_2 can be used to define lower confidence limits on $POD(a)$ which are valid for much smaller POD data sets than previously possible. However, as sample size decreases, the lower confidence limit curve becomes increasingly conservative with respect to the best estimate curve. Whether such confidence limit curves will be practically useful will depend on the available data and the requirements of the particular application.

Annex C – REPRODUCIBILITY AND REPEATABILITY IN EDDY CURRENT TESTS

C.1 INTRODUCTION

Following many years of research it has still proven difficult to determine an accurate probability of detection (POD) for non-destructive inspection (NDI) operations. The RTO/AGARD concept is a NATO attempt to define a systematic approach to solving the POD problem from research work carried out over many years. The Applied Vehicle Technology (AVT) Panel 051 Working Group was to embark on a 3-year programme with the aim of producing a realistic procedure and international database of NDI results. The NDI input would be from a user perspective to the group, which is mainly comprised of scientific and statistical experts.

C.2 IN-HOUSE TRIAL

A small in-house trial was set up to attempt to illustrate the difference in probe handling, equipment set-up and interpretation between experienced NDI operators, and also to look into the similarity and any linearity in the results between like probes. The trial focused on 3 individual eddy current methods using our own in-service eddy current equipment – hand scanning with the Hocking Locator UH meter display (Figure C-1), hand scanning with the Hocking Locator 2 impedance plane (Figure C-2) and rotary eddy current with the Rohmann Rototest (Figure C-3).



Figure C-1: A Photograph of the Hocking Locator UH Instrument.



Figure C-2: A Photograph of the Hocking Locator 2 Instrument.

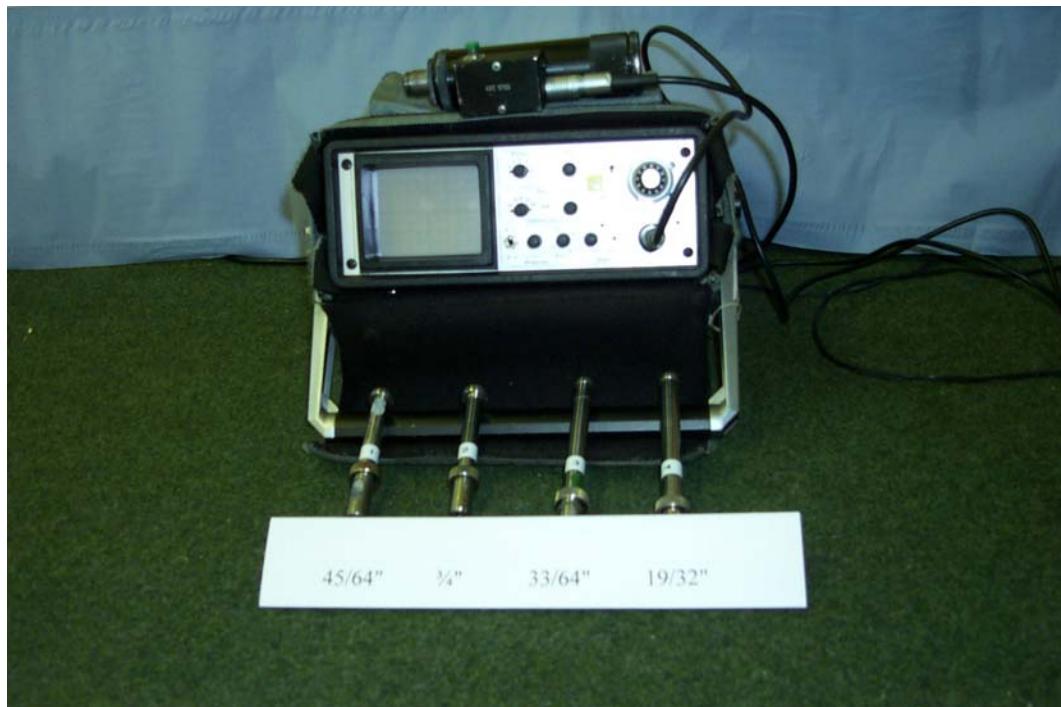


Figure C-3: A Photograph of the Rohmann Rototest Instrument.

NDI technicians used for the trial all had a minimum of 4 years experience and each operator performed the trial anonymously. The procedure for each method was clearly stated to ensure repeatability between operators.

For the hand-held method, 3 sets of six probes were used. All the probes were unused prior to the trial and were individually numbered (Figure C-4).



Figure C-4: A Photograph of the Different Probes used in this Trial.

C.3 EDDY CURRENT TRIAL – METER INSTRUMENT SET-UP

The trial was to determine the standard of probe set-up between NDI operators for a variety of hand-held eddy current probes. It also established if probes are uniform in their performance characteristics.

In the United Kingdom, the Hocking Locator UH has been the in-service general purpose eddy current instrument for at least 10 years so; all the operators were well established with its operation.

The operators were given 2 eddy current reference blocks numbered 1 (new block) and 2 (used block), each having 0.2, 0.5 and 1.0 mm depth, 0.1 mm width spark-eroded slots cut their full width. They were requested to set-up each probe to give a 50% screen deflection from the 0.5 mm slot on block 1 and then record the gain level to achieve this on the results table. On the same block, they were then requested to record the average needle deflection taken from 3 passes over the 0.2 and the 1.0 mm slots and record these on the results table. The final request was to record the needle deflection from the 0.5 mm slot in block number 2. This was then repeated with each probe in the set, and then repeated with the other 2 sets of probes.

Five (5) operators carried out the trial with the Locator UH. The tables below are their results (Tables C-1, C-2 and C-3). *NOTE: The results have been reproduced in chart format at the end of this report.*

Table C-1: Probe Set 1

Operator		1/1	1/2	1/3	1/4	1/5	1/6
1	Gain at set up	196	200	214	212	229	193
	% Needle Swing 0.2 mm Slot	22	22	24	22	22	22
	% Needle Swing 1.0 mm Slot	66	68	70	64	68	67
	% Needle Swing 0.5 mm Slot (Block 2)	46	46	48	46	45	46
2	Gain at set up	191	205	198	184	219	181
	% Needle Swing 0.2 mm Slot	16	20	24	30	18	20
	% Needle Swing 1.0 mm Slot	82	72	85	76	76	70
	% Needle Swing 0.5 mm Slot (Block 2)	60	50	54	50	56	54
3	Gain at set up	202	223	196	191	242	173
	% Needle Swing 0.2 mm Slot	18	18	22	19	20	21
	% Needle Swing 1.0 mm Slot	58	64	60	61	65	64
	% Needle Swing 0.5 mm Slot (Block 2)	42	48	49	45	45	45
4	Gain at set up	195	209	200	196	222	199
	% Needle Swing 0.2 mm Slot	18	19	20	18	18	18
	% Needle Swing 1.0 mm Slot	64	64	62	63	62	63
	% Needle Swing 0.5 mm Slot (Block 2)	48	48	47	48	47	47
5	Gain at set up	193	213	194	225	235	164
	% Needle Swing 0.2 mm Slot	22	22	24	27	22	24
	% Needle Swing 1.0 mm Slot	70	71	64	69	70	65
	% Needle Swing 0.5 mm Slot (Block 2)	49	52	50	52	50	45

Table C-2: Probe Set 2

Operator		2/1	2/2	2/3	2/4	2/5	2/6
1	Gain at set up	258	258	256	250	243	290
	% Needle Swing 0.2 mm Slot	19	18	19	18	20	18
	% Needle Swing 1.0 mm Slot	85	85	86	86	91	86
	% Needle Swing 0.5 mm Slot (Block 2)	49	46	48	51	51	46
2	Gain at set up	246	250	246	242	219	275
	% Needle Swing 0.2 mm Slot	20	18	20	20	20	18
	% Needle Swing 1.0 mm Slot	90	84	90	86	88	90
	% Needle Swing 0.5 mm Slot (Block 2)	50	50	52	50	52	54
3	Gain at set up	249	264	250	236	216	286
	% Needle Swing 0.2 mm Slot	18	19	18	18	18	18
	% Needle Swing 1.0 mm Slot	85	87	86	85	83	85
	% Needle Swing 0.5 mm Slot (Block 2)	49	49	52	49	47	49
4	Gain at set up	236	265	264	238	238	228
	% Needle Swing 0.2 mm Slot	18	17	18	18	18	20
	% Needle Swing 1.0 mm Slot	85	85	85	85	88	87
	% Needle Swing 0.5 mm Slot (Block 2)	48	50	50	49	50	50
5	Gain at set up	259	257	252	241	229	296
	% Needle Swing 0.2 mm Slot	19	19	19	18	19	18
	% Needle Swing 1.0 mm Slot	87	83	86	84	86	88
	% Needle Swing 0.5 mm Slot (Block 2)	50	49	50	51	51	48

Table C-3: Probe Set 3

Operator		3/1	3/2	3/3	3/4	3/5	3/6
1	Gain at set up	266	208	221	221	251	196
	% Needle Swing 0.2 mm Slot	22	22	22	22	21	20
	% Needle Swing 1.0 mm Slot	66	65	68	65	71	68
	% Needle Swing 0.5 mm Slot (Block 2)	45	47	46	44	48	46
2	Gain at set up	223	189	189	196	195	185
	% Needle Swing 0.2 mm Slot	30	32	30	26	30	28
	% Needle Swing 1.0 mm Slot	80	90	74	72	84	78
	% Needle Swing 0.5 mm Slot (Block 2)	60	60	60	50	56	60
3	Gain at set up	245	202	212	219	233	200
	% Needle Swing 0.2 mm Slot	22	22	23	22	22	21
	% Needle Swing 1.0 mm Slot	66	67	68	68	68	67
	% Needle Swing 0.5 mm Slot (Block 2)	47	46	52	51	48	49
4	Gain at set up	227	191	205	226	176	191
	% Needle Swing 0.2 mm Slot	20	18	21	22	20	20
	% Needle Swing 1.0 mm Slot	66	68	70	72	74	68
	% Needle Swing 0.5 mm Slot (Block 2)	48	48	51	51	49	49
5	Gain at set up	246	235	227	227	243	219
	% Needle Swing 0.2 mm Slot	23	22	26	23	20	24
	% Needle Swing 1.0 mm Slot	69	68	69	67	69	71
	% Needle Swing 0.5 mm Slot (Block 2)	50	58	52	50	52	54

C.4 EDDY CURRENT TRIAL – IMPEDANCE PLANE SET-UP

This trial also determined the standard of probe set-up between experienced NDI operators for the same hand-held eddy current probes as used for the meter instrument. Although experienced with Impedance Plane eddy current testing, the Hocking Locator 2 instrument was new to the operators at the time of the trial.

The procedure for the trial was similar to the meter instrument. For the initial set-up, the alarm gate was set at 75 screen height and then the gain level required to break this alarm while scanning the 0.5 mm slot on Block 1 was recorded. For the 0.2 and 1.0 mm slots the alarm gate level was recorded when it just touched the screen indication.

The tables below show the results (Tables C-4, C-5 and C-6).

Table C-4: Probe Set 1

Operator		1/1	1/2	1/3	1/4	1/5	1/6
1	Gain at set up (dB)	35.9	36.2	33.2	35.3	39.5	34.7
	Top Alarm Gate – 0.2 mm Slot	35	33	36	35	35	36
	Top Alarm Gate – 1.0 mm Slot	108	104	111	105	113	103
	Top Alarm Gate – 0.5 mm Slot (Block 2)	72	74	78	76	75	70
2	Gain at set up (dB)	37.6	37.3	32.7	36.7	40.8	34.9
	Top Alarm Gate – 0.2 mm Slot	34	33	38	35	34	31
	Top Alarm Gate – 1.0 mm Slot	108	106	107	104	103	103
	Top Alarm Gate – 0.5 mm Slot (Block 2)	71	68	81	75	72	71
3	Gain at set up (dB)	37.5	37.7	34.1	36.5	40.7	34.1
	Top Alarm Gate – 0.2 mm Slot	33	34	34	35	35	35
	Top Alarm Gate – 1.0 mm Slot	105	108	104	102	107	105
	Top Alarm Gate – 0.5 mm Slot (Block 2)	71	74	73	73	75	75
4	Gain at set up (dB)	38.9	37.4	32.8	36	39.2	32.4
	Top Alarm Gate – 0.2 mm Slot	32	35	36	33	34	36
	Top Alarm Gate – 1.0 mm Slot	106	112	108	99	104	101
	Top Alarm Gate – 0.5 mm Slot (Block 2)	73	75	76	70	72	72
5	Gain at set up (dB)	40	37.3	42.1	36.3	38	36
	Top Alarm Gate – 0.2 mm Slot	38	33	30	34	33	32
	Top Alarm Gate – 1.0 mm Slot	115	109	111	117	114	116
	Top Alarm Gate – 0.5 mm Slot (Block 2)	76	78	74	72	70	76

Table C-5: Probe Set 2

Operator		2/1	2/2	2/3	2/4	2/5	2/6
1	Gain at set up (dB)	40.7	41.8	40.1	39.5	39.1	46.2
	Top Alarm Gate – 0.2 mm Slot	29	25	28	28	28	28
	Top Alarm Gate – 1.0 mm Slot	148	125	135	122	138	133
	Top Alarm Gate – 0.5 mm Slot (Block 2)	78	67	77	68	74	75
2	Gain at set up (dB)	40.9	40.5	40.6	40.1	39.2	42.3
	Top Alarm Gate – 0.2 mm Slot	29	29	30	29	29	28
	Top Alarm Gate – 1.0 mm Slot	139	137	141	134	134	139
	Top Alarm Gate – 0.5 mm Slot (Block 2)	72	74	78	74	74	76
3	Gain at set up (dB)	41.3	40.7	40.5	40.2	40	42.5
	Top Alarm Gate – 0.2 mm Slot	28	27	26	27	27	27
	Top Alarm Gate – 1.0 mm Slot	137	132	133	134	131	131
	Top Alarm Gate – 0.5 mm Slot (Block 2)	75	77	74	72	74	74
4	Gain at set up (dB)	34.3	34.7	35.6	32.9	33.3	33.6
	Top Alarm Gate – 0.2 mm Slot	35	33	31	34	35	35
	Top Alarm Gate – 1.0 mm Slot	102	104	107	102	105	111
	Top Alarm Gate – 0.5 mm Slot (Block 2)	72	73	68	74	72	73
5	Gain at set up (dB)	37.8	40.5	41.9	39.8	42.2	41.1
	Top Alarm Gate – 0.2 mm Slot	23	29	27	27	28	26
	Top Alarm Gate – 1.0 mm Slot	121	139	141	127	132	134
	Top Alarm Gate – 0.5 mm Slot (Block 2)	67	69	65	66	67	69

Table C-6: Probe Set 3

Operator		3/1	3/2	3/3	3/4	3/5	3/6
1	Gain at set up (dB)	31.1	31.4	33	33.1	31.9	32.3
	Top Alarm Gate – 0.2 mm Slot	36	32	36	35	36	33
	Top Alarm Gate – 1.0 mm Slot	100	104	104	106	106	106
	Top Alarm Gate – 0.5 mm Slot (Block 2)	75	75	75	77	76	74
2	Gain at set up (dB)	34.3	32.7	33.5	33.7	34.4	33.8
	Top Alarm Gate – 0.2 mm Slot	34	34	34	36	34	33
	Top Alarm Gate – 1.0 mm Slot	109	107	103	109	109	111
	Top Alarm Gate – 0.5 mm Slot (Block 2)	73	76	72	77	76	76
3	Gain at set up (dB)	34.7	33.8	33.2	34.4	36	33
	Top Alarm Gate – 0.2 mm Slot	34	33	33	40	33	34
	Top Alarm Gate – 1.0 mm Slot	101	106	100	106	109	108
	Top Alarm Gate – 0.5 mm Slot (Block 2)	72	73	72	82	74	75
4	Gain at set up (dB)	41.4	40.9	40.6	41.2	41.2	43.2
	Top Alarm Gate – 0.2 mm Slot	27	27	26	28	28	27
	Top Alarm Gate – 1.0 mm Slot	137	132	130	128	150	141
	Top Alarm Gate – 0.5 mm Slot (Block 2)	66	77	68	61	84	74
5	Gain at set up (dB)	34.2	37.6	34.8	39.6	39.1	32.8
	Top Alarm Gate – 0.2 mm Slot	37	45	47	41	43	47
	Top Alarm Gate – 1.0 mm Slot	100	115	110	112	111	108
	Top Alarm Gate – 0.5 mm Slot (Block 2)	68	72	66	69	71	70

The above results have been illustrated in chart format at the end of this report.

C.5 EDDY CURRENT TRIAL – ROTARY

This trial is to determine the standard of probe set-up between NDI operators for four rotary eddy current probes of diameters 45/64", 3/4", 33/64" and 19/32".

The standard set-up for the Rototest equipment is to obtain a vertical screen signal of 40% screen height from the 1/16" cross-drilled hole in the reference block. This point on the screen is one main-scale division plus three sub-divisions (Figure C-5). From this setting the operator would then add 10dB, which would take the signal off screen. The phase angle of the signal would then be adjusted to 30° from vertical to be able to distinguish between fault indications and mechanical damage.

ROTARY EDDY CURRENT SCREEN

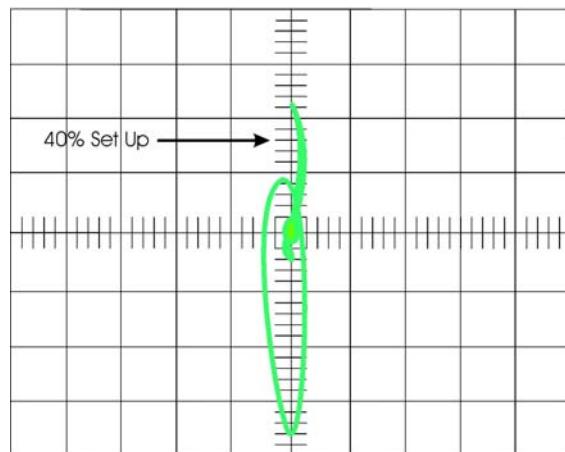


Figure C-5: The Rototest Instrument Response for the Calibration Test.

For this trial the operator was requested to first record the gain setting (dB) at the initial 40% screen height set-up and then only add 5dB to keep the signal on screen.

The rotary reference block used in the trial had 5 spark-eroded slots cut into the bore of the relevant size holes to match the probe sizes. These slots were numbered 1-5, with number 1 and 5 being bore edge slots (1 at the rear – 5 nearest the operator). The slots numbered 2, 3 and 4 were aligned down the bore and were cut at 1, 0.5 and 0.2 mm depths (about 3 mm length) respectively; depths as the slots on the standard hand-held eddy current probe reference blocks.

The operators in the trial were requested to record the percentage screen height for each indication from the slots in the bore for each size of probe. For this trial, 9 experienced operators were used and their results are reproduced below (Table C-7). It appears that operator No 8 has misread the instructions as the results for Slots 2, 3 and 4 for probes 1, 2 and 3 are ascending rather than descending, as would be expected.

Table C-7: The Percentage Screen Height from the Slots in the Bore

Operator		Probe 1 45/64"	Probe 2 3/4"	Probe 3 33/64"	Probe 4 19/32"
1	Gain at Set-Up (dB)	23	19	22	13
	Slot 1 – % Screen Height	45	75	50	70
	Slot 2 – % Screen Height	50	70	75	70
	Slot 3 – % Screen Height	30	15	45	50
	Slot 4 – % Screen Height	10	10	25	25
	Slot 5 – % Screen Height	45	90	50	70
2	Gain at Set-Up (dB)	30	20	28	14
	Slot 1 – % Screen Height	65	10	20	65
	Slot 2 – % Screen Height	20	40	75	25
	Slot 3 – % Screen Height	65	70	75	50
	Slot 4 – % Screen Height	75	40	75	70
	Slot 5 – % Screen Height	65	30	100	65

ANNEX C – REPRODUCIBILITY AND
REPEATABILITY IN EDDY CURRENT TESTS

Operator		Probe 1 45/64"	Probe 2 3/4"	Probe 3 33/64"	Probe 4 19/32"
3	Gain at Set-Up (dB)	30	25	26	17
	Slot 1 – % Screen Height	60	60	80	55
	Slot 2 – % Screen Height	75	75	80	70
	Slot 3 – % Screen Height	55	55	60	45
	Slot 4 – % Screen Height	20	25	25	25
	Slot 5 – % Screen Height	70	85	60	55
4	Gain at Set-Up (dB)	29.5	24.5	25	16
	Slot 1 – % Screen Height	65	50	70	50
	Slot 2 – % Screen Height	80	70	80	65
	Slot 3 – % Screen Height	65	45	60	45
	Slot 4 – % Screen Height	20	25	25	25
	Slot 5 – % Screen Height	70	80	60	45
5	Gain at Set-Up (dB)	35	30	34	24
	Slot 1 – % Screen Height	65	55	85	85
	Slot 2 – % Screen Height	55	60	95	100
	Slot 3 – % Screen Height	40	45	60	70
	Slot 4 – % Screen Height	15	20	20	30
	Slot 5 – % Screen Height	50	75	55	75
6	Gain at Set-Up (dB)	30	22	26	17
	Slot 1 – % Screen Height	67	42	65	75
	Slot 2 – % Screen Height	50	37	75	75
	Slot 3 – % Screen Height	37	25	37	65
	Slot 4 – % Screen Height	12	15	25	25
	Slot 5 – % Screen Height	50	37	50	75
7	Gain at Set-Up (dB)	29	19	28	14
	Slot 1 – % Screen Height	65	60	55	65
	Slot 2 – % Screen Height	80	85	75	80
	Slot 3 – % Screen Height	65	60	55	65
	Slot 4 – % Screen Height	25	20	25	30
	Slot 5 – % Screen Height	60	90	50	55
8	Gain at Set-Up (dB)	30	23	29	19
	Slot 1 – % Screen Height	50	60	45	85
	Slot 2 – % Screen Height	15	15	15	40
	Slot 3 – % Screen Height	55	40	35	90
	Slot 4 – % Screen Height	75	55	50	45
	Slot 5 – % Screen Height	50	40	55	85
9	Gain at Set-Up (dB)	30	28	30	20
	Slot 1 – % Screen Height	70	95	100	85
	Slot 2 – % Screen Height	75	90	100	100
	Slot 3 – % Screen Height	55	65	70	80
	Slot 4 – % Screen Height	20	20	30	40
	Slot 5 – % Screen Height	60	80	75	85

C.6 CONCLUSION

The small trial has highlighted that even with the most experienced NDI operators working in ideal conditions, there is a considerable difference in the results. This can be accounted to probe handling, variance between probe, signal interpretation and human error (not reading laid down instructions!).

Even the Meter Instrument, which was equipment that was well known and has been regularly used since the late eighties, produced quite a large range of results. The Impedance Plane equipment at the time of the trial was relatively new, however the results were encouraging.

It was perplexing that the rotary trial produced such a large range of results, which could well be attributed to probe handling. The varying pressure between operators of the probe against the 1/16" cross-drilled hole at instrument set-up, and against the simulated fault during inspection, will affect the results considerably. With this instrument, it is not possible to select full persistence of the screen, so the probe has to be held steady at the maximum signal from the simulated faults prior to taking a reading.

Overall, the trial has emphasised that between experienced NDI Technicians, operating in ideal conditions, there can be quite a variance in the results. This discrepancy can only be amplified when inspecting in difficult access situations, inadequately prepared areas, in inclement weather, being pressurised, or any of the many situations an NDI technician could find themselves under whilst trying to carry out an NDI technique.

C.7 RESULTS CHARTS – METER DISPLAY INSTRUMENT

The length of the red indicators represents the maximum and minimum values for each probe at set-up (Figure C-6), as recorded by the operators in the trial.

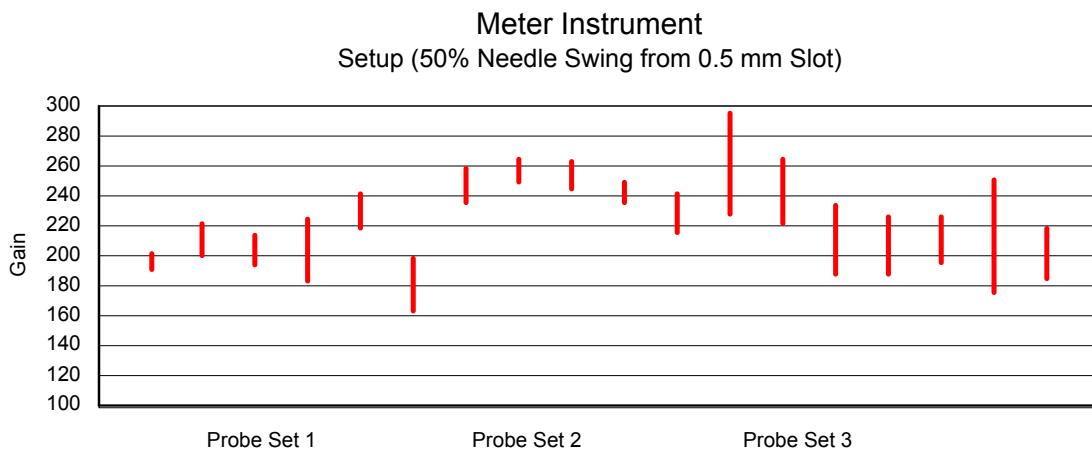


Figure C-6: A Plot of the Variation Required to Achieve a 50% Needle Swing from a 0.5 mm Slot.

Here and for the 2 probe sets in the charts below (Figure C-7 and C-8), the maximum and minimum variation for needle swing is indicated – probes 1/1 and 1/3 having the greatest range from the 1.0 mm slot.

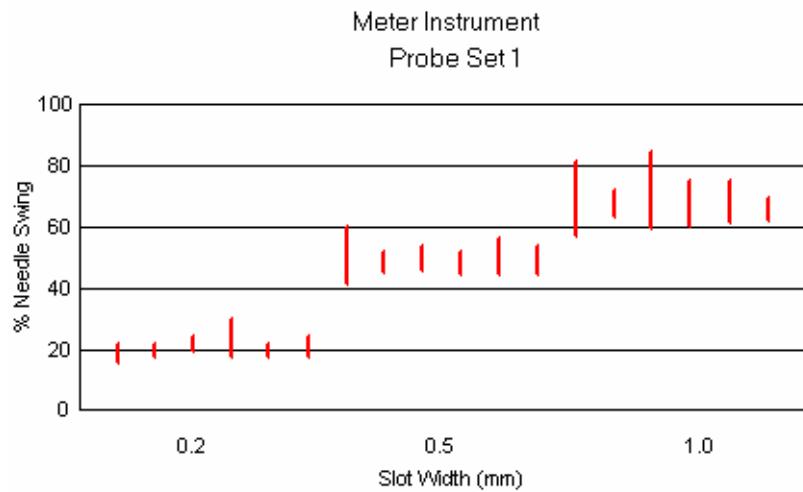


Figure C-7: A Plot of the Variation in Needle Swing for Nominally Equivalent Probes from Set 1, for Different Slot Widths.

For Probe Set 2 it was encouraging to note the small range in the results (Figure C-8). This probe set used the unshielded probe which would be less susceptible to poor probe handling than the shielded probes. It is interesting to note that the majority of probes used for in-service techniques are shielded.

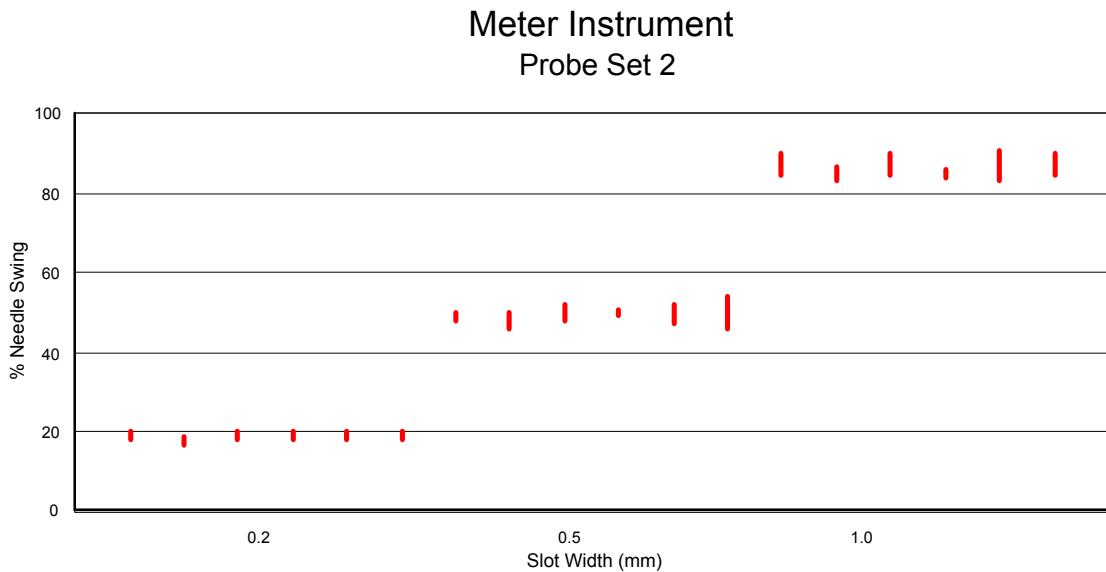


Figure C-8: A Plot of the Variation in Needle Swing for Nominally Equivalent Probes from Set 2, for Different Slot Widths.

Probe Set 3 had the largest range of results for the 3 slots (Figure C-9). This was a shielded pencil-probe which should be able to be kept in the correct vertical position during scanning and therefore would expect to be better than the results from Probe Set 1.

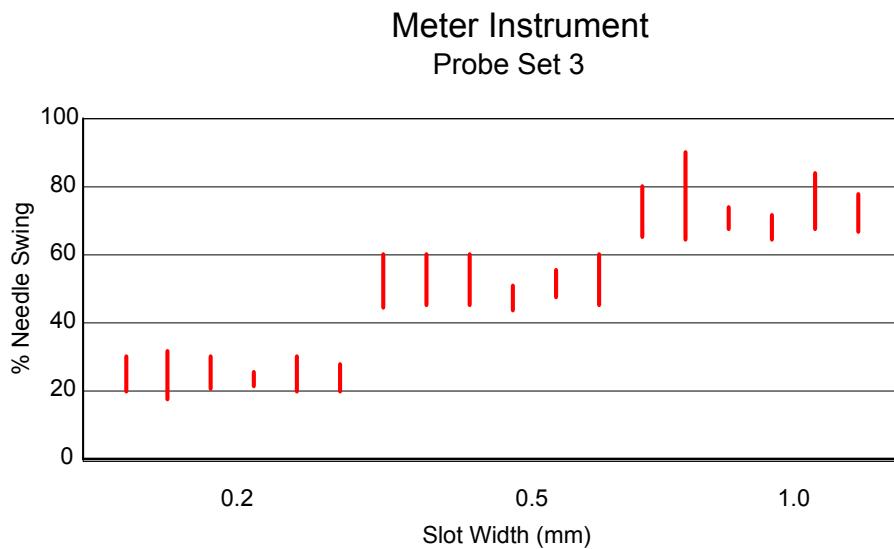


Figure C-9: A Plot of the Variation in Needle Swing for Nominally Equivalent Probes from Set 3, for Different Slot Widths.

C.8 RESULTS CHARTS – IMPEDANCE PLANE DISPLAY EDDY CURRENT

From the set-up of the probes, Probe Set 1 (Figure C-10) had the better range, which has been reflected in the results in the chart. This, in comparison to the other results, is a vast improvement.

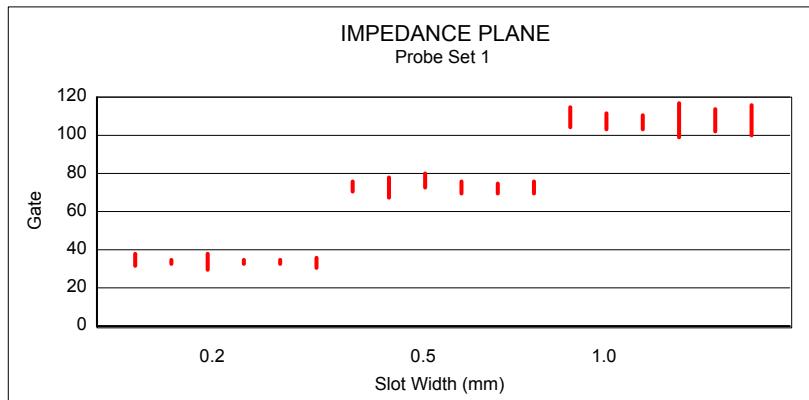


Figure C-10: A Plot of the Variation in Amplitude for Nominally Equivalent Probes from Set 1, for Different Slot Widths.

It is interesting to note that the range for the 1.0 mm slot in both Probe Set 2 (Figure C-11) and Set 3 (Figure C-12) are significantly greater than those from the other 2 slots. Figure C-13 compares the performance of the three different probe sets, plotting the variation in gain required to achieve the same signal on a 0.5 mm slot for all the different probes.

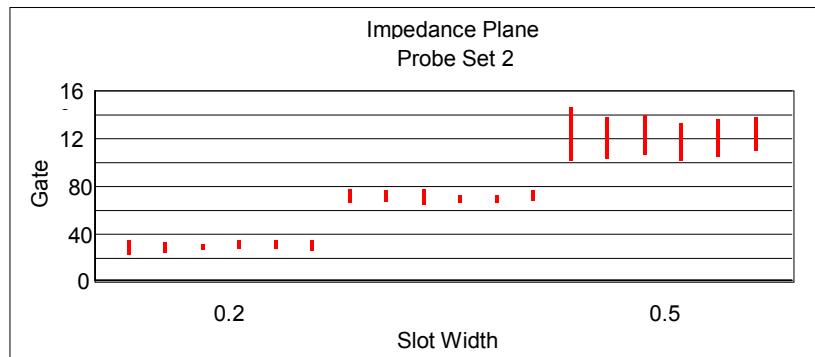


Figure C-11: A Plot of the Variation in Amplitude for Nominally Equivalent Probes from Set 2, for Different Slot Widths.

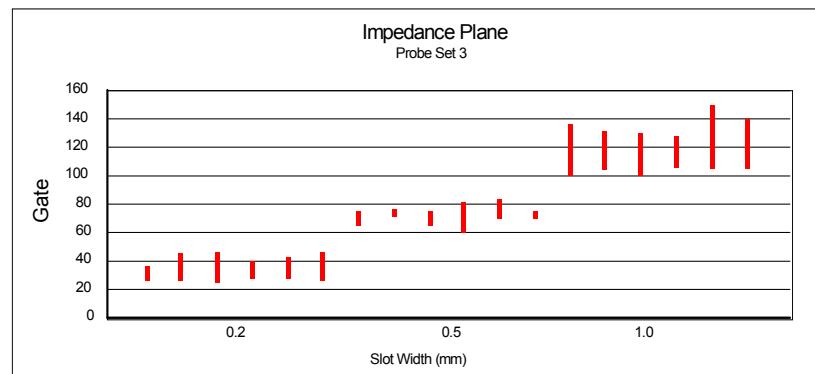


Figure C-12: A Plot of the Variation in Amplitude for Nominally Equivalent Probes from Set 3, for Different Slot Widths.

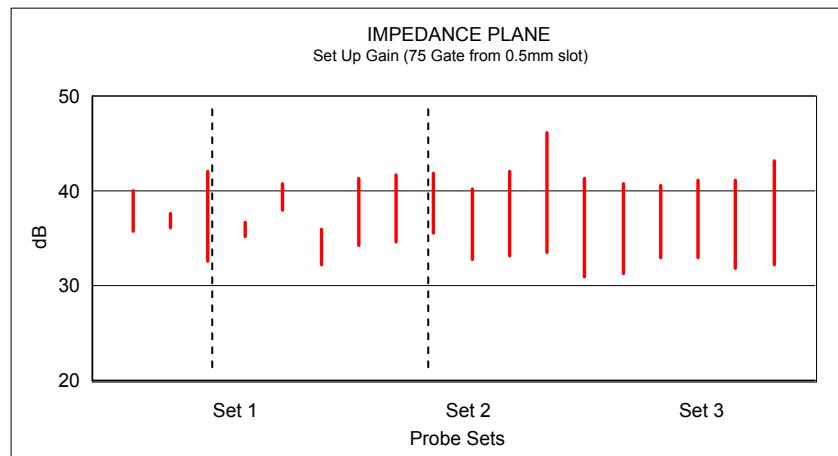


Figure C-13: A Plot of the Variation in Gain Required to Achieve the Same Signal from a 0.5 mm Slot, using Different Probes.

C.9 RESULTS CHARTS – ROTARY EDDY CURRENT

To chart the results from the rotary inspection, it was decided to remove the highest and lowest figures (see Table C-7). However, all probe sets have revealed a considerably large variation. Probe 1 (Figure C-14) has the better set of results of the 4 probes.

NOTE: The faults at 1 and 5 are the edge slots and 2, 3 and 4 are the 1.0, 0.5 and 0.2 mm slots, respectively.

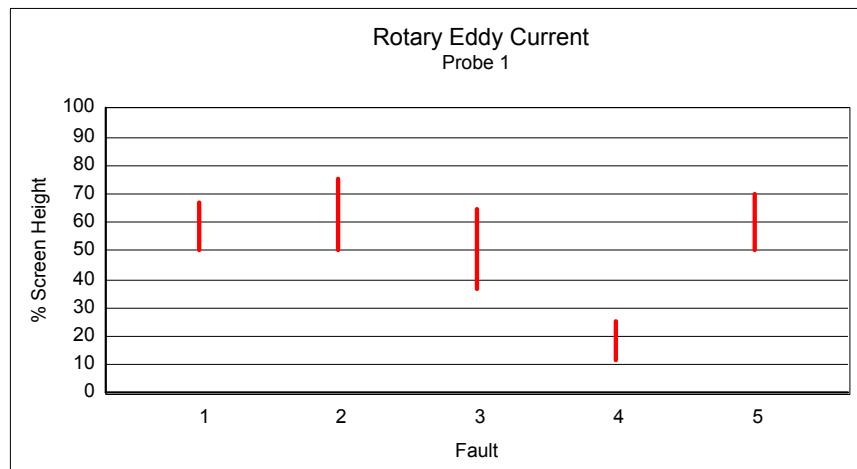


Figure C-14: A Plot of the Variation in Signal Amplitude Recorded by Different Inspectors using the Same Probe (Probe 1) on Five Different Slots.

For Probe 2 (Figure C-15), the 1.0 mm slot appears to be the most difficult for screen height identification with operators results varying from 37 to 85% screen height.

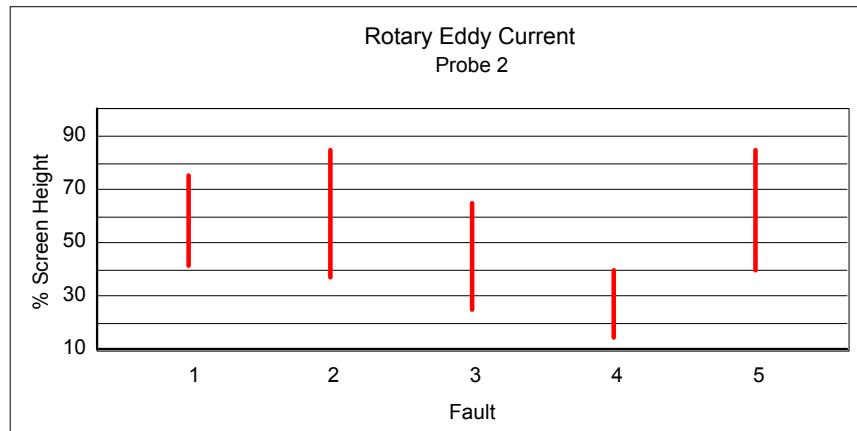


Figure C-15: A Plot of the Variation in Signal Amplitude Recorded by Different Inspectors using the Same Probe (Probe 2) on Five Different Slots.

Figure C-16 plots the variation in signal which was obtained by different inspectors, using the same probe, on faults 1 through 5.

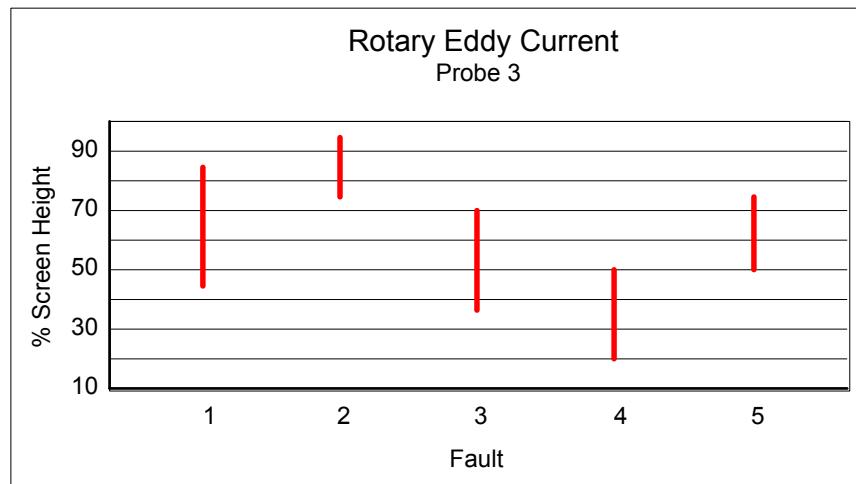


Figure C-16: A Plot of the Variation in Signal Amplitude Recorded by Different Inspectors using the Same Probe (Probe 3) on Five Different Slots.

For Probe 4 (Figure C-17), the range of percentage screen height results would be expected to descend from faults 2 to 4, with 4 being the smallest slot. Faults 1 and 5 (the edge slots) should be a similar size. With number 5 being closer to the operator, the range (as in this chart) should be smaller. Although 1 and 5 are similar ranges for Probe 4 as expected, the fault 2 range of results raises concern, being from 40% to 100%.

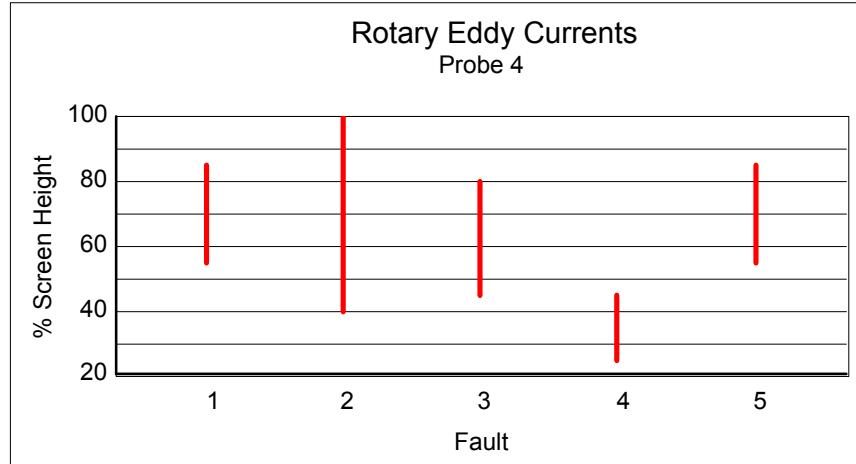


Figure C-17: A Plot of the Variation in Signal Amplitude Recorded by Different Inspectors using the Same Probe (Probe 4) on Five Different Slots.

The set-up gain chart of Figure C-18 also has a considerable range, which would subsequently contribute to the range of results during the trial. The 5dB gain with this equipment will increase the signal by 50%.

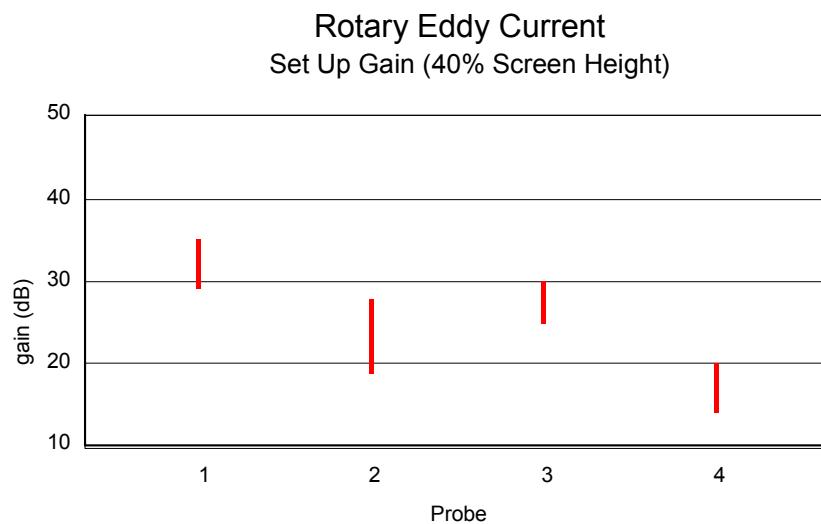


Figure C-18: A Plot of the Variations in Gain Required by Different Inspectors to Achieve a Signal of 40% Screen Height on the Same Calibration Slot.

Annex D – CUMULATIVE DISTRIBUTION FUNCTION (CDF) OF DETECTED CRACKS

D.1 INTRODUCTION

To illustrate the PDF/CDF approach mentioned in Section 5.4 of the main report, the inspection data of an AGARD round-robin NDI demonstration programme and the in-service inspection data of a control point of the F-16 airframe structure have been reviewed (see Fahr et al. (1995) and Heida and Grooteman (1998), respectively).

D.2 AGARD ROUND-ROBIN NDI DEMONSTRATION PROGRAMME

Fahr et al. (1995) give the results of an AGARD round-robin NDI demonstration programme in which six laboratories from four NATO countries participated. In this programme, several NDI procedures were evaluated for the detection of low cycle fatigue cracks in the bolt holes of service-expired compressor disks and spacers of the J85-CAN40 engine. The material of the components was precipitation hardened martensitic stainless steel (AM355). The NDI procedures included manual and (semi)-automated eddy current, automated ultrasonics, X-ray, optical microscopy, liquid penetrant and magnetic particle inspection. After inspection, the components were destructively examined for the verification and sizing of cracks. The database of Fahr et al. (1995) comprises a large amount of “hit” data, “miss” data and false calls for a total of seven compressor disks and six spacers inspected with the NDI techniques mentioned. Finally, POD and lower 95% confidence curves as functions of crack size were determined.

Figure D-1 gives an example of the PDF/CDF approach with a plot of the CDF-hits curve and mean POD curve (50% confidence level) of the manual eddy current inspection results from Fahr et al. (1995). The CDF curve was drawn based on the 79 “hit” data only. The POD curve was constructed from 79 “hit” and 206 “miss” data. A log-normal distribution function was assumed for the curves. The location (μ) and scale (σ) parameters were determined with the least-squares method (CDF curve) or with the MLE method (POD curve), resulting in (μ, σ) values of $(2.3, 1.2)$ mm and $(1.6, 0.7)$ mm, respectively.

Figure D-1 shows that the CDF-hits curve is located to the right of the mean POD curve, i.e. it is conservative. An arbitrary 90% probability criterion yields the crack lengths of 3.8 mm and 2.4 mm for the CDF and POD curve, respectively. It is emphasised that these values cannot be compared directly: 2.4 mm is the crack length for which there is a 90% probability of detection (confidence level 50%), while 3.8 mm is the crack length for which there is a 90% probability that the detected cracks have a length less than or equal to 3.8 mm. For this inspection case, the CDF-hits curve gives a conservative estimate of the reliably detectable crack length a_d , here arbitrarily defined as the crack length for which there is a mean POD of 90%.

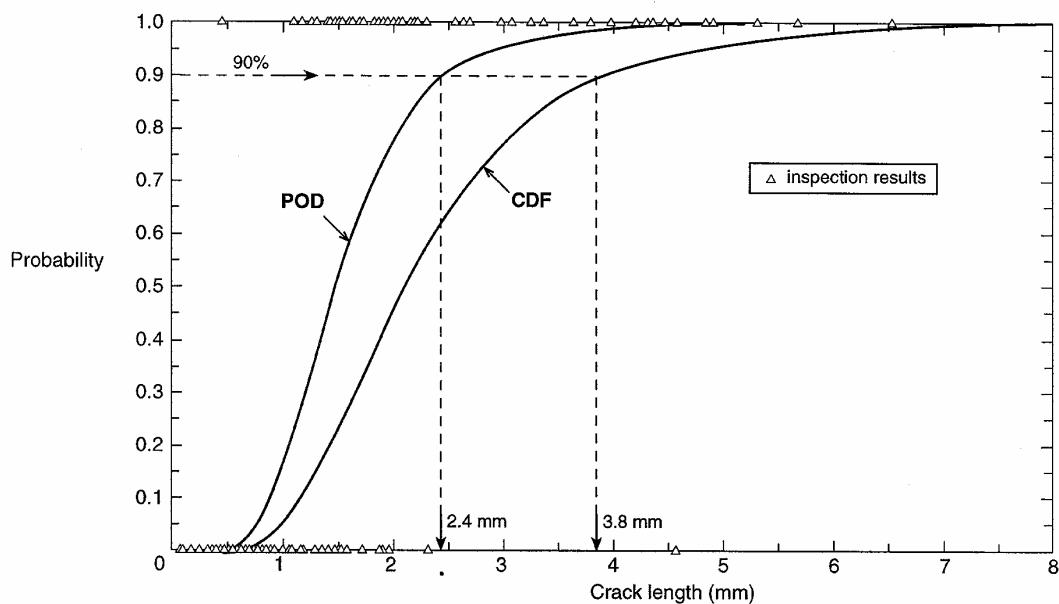


Figure D-1: Mean POD Curve for the “Hit/Miss” Data and CDF-Hits Curve for the “Hit” Data of the Manual Eddy Current Inspection Database of Fahr et al. (1995).

The AGARD round-robin NDI demonstration programme resulted in eighteen data sets for the NDI techniques investigated. A further comparison between POD and CDF curves was performed using seven other data sets from Fahr et al. (1995), viz. one data set for liquid penetrant inspection, two data sets for magnetic particle inspection and four data sets for (semi)-automated eddy current inspection. For all inspection cases, the CDF-hits curve is located to the right of the mean POD curve, i.e. it is conservative. In addition to the CDF curves for the “hit” data, CDF curves for the “miss” data were calculated also assuming a log-normal distribution function. As can be expected, these CDF-misses curves were all located to the left of the mean POD curve. A remarkable observation was that the goodness-of-fit for the CDF-misses curves is much better than that for the CDF-hits curves. This is illustrated in Figures D-2 and D-3 for the “hit” and “miss” data of the manual eddy current inspection results of Fahr et al. (1995).

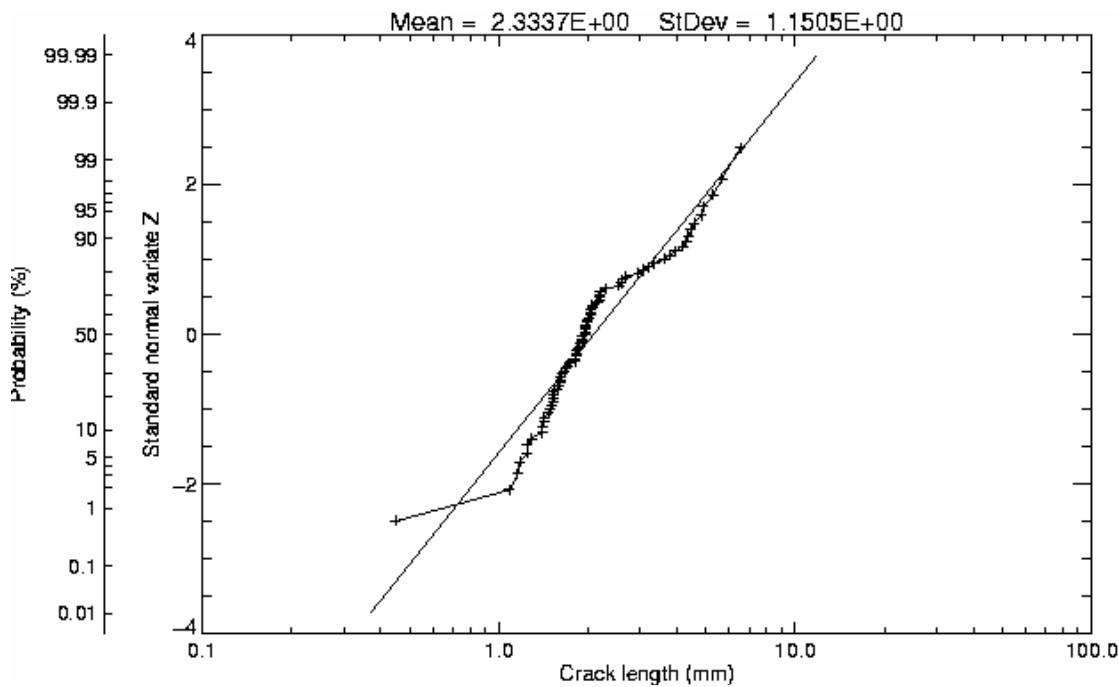


Figure D-2: Goodness-of-Fit for the Log-Normal PDF Estimation for the “Hit” Data of the Manual Eddy Current Inspection Results of Fahr et al. (1995). Standard normal variate $z = (\ln(a) - \mu_y) / \sigma_y$ and its corresponding cumulative probability versus the crack length detected, plotted on log-normal probability paper.

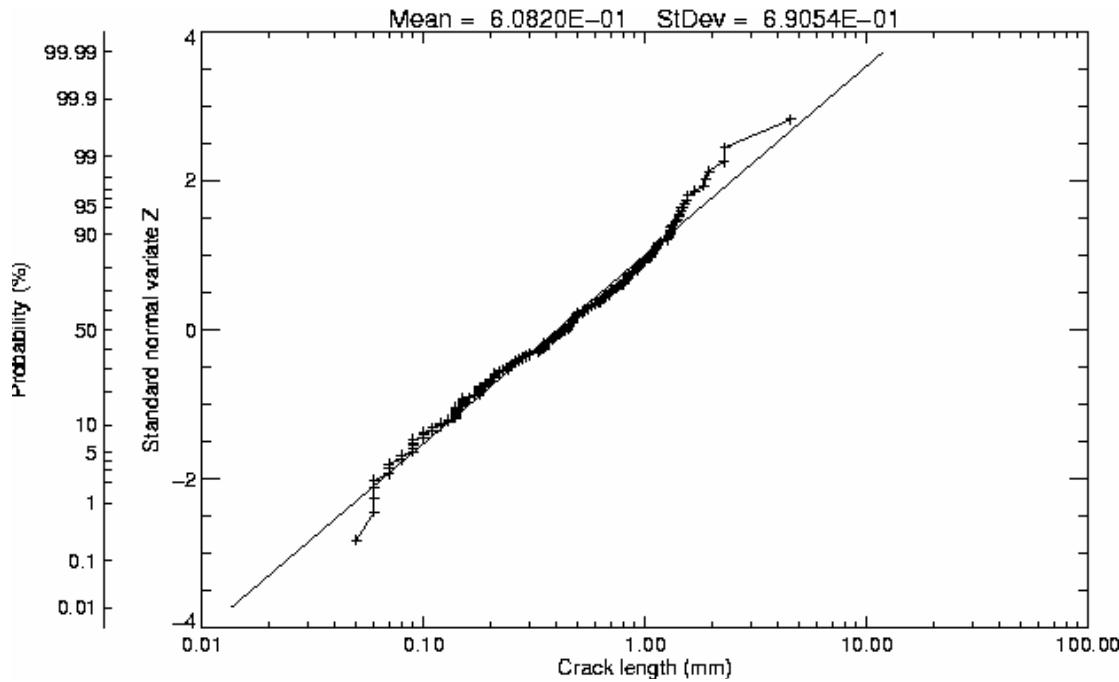


Figure D-3: Goodness-of-Fit for the Log-Normal PDF Estimation for the “Miss” Data of the Manual Eddy Current Inspection Results of Fahr et al. (1995). Standard normal variate $z = (\ln(a) - \mu_y) / \sigma_y$ and its corresponding cumulative probability versus the crack length missed, plotted on log-normal probability paper.

A rough comparison between the different POD and CDF curves is made in Table D-1, which gives the crack lengths for which there is a 90% probability value. The table shows for all eight inspection cases that the 90% CDF-hits values are higher than the 90% POD values. It can be concluded that for this inspection configuration, the CDF-hits curve gives a conservative estimate of the detectable crack length.

**Table D-1: Comparison of POD, CDF-Hits and CDF-Misses Curves using the Data of an AGARD Round-Robin NDI Demonstration Programme, Fahr et al. (1995).
Crack lengths (in mm) for which there is a 90% probability value.**

Inspection Technique	POD	CDF-hits	CDF-misses
LPI (I)	2.4	4.1	0.9
MPI (I)	3.3	4.8	1.4
MPI (II)	1.8	3.8	0.7
ECI-M (I)	2.4	3.8	1.3
ECI (III)	0.8	2.9	0.4
ECI (V)	0.9	5.0	0.8
ECI (VI)	0.7	3.1	0.4
ECI-A (IV)	1.2	2.8	0.6

POD – Probability of Detection

CDF – Cumulative Distribution Function

Inspection Technique:

- LPI: Liquid Penetrant Inspection
- MPI: Magnetic Particle Inspection
- ECI-M: Manual Eddy Current Inspection
- ECI: Semi-Automated Eddy Current Inspection
- ECI-A: Automated Eddy Current Inspection

D.3 F-16 FUSELAGE LONGERON TAB RADII

Heida and Grooteman (1998) give an evaluation of the in-service inspection data of a control point of the F-16 airframe structure. The database, status March 1998, comprises 28 “hit” and 36 “miss” data points back-extrapolated using a durability crack growth curve. The corresponding CDF-hits curve and POD curve have been discussed in Chapter 5.4.

An update of the inspection database, status May 2000, will be discussed in the following section.

D.3.1 General Data

a) Part

The F-16 centre fuselage longeron is a tee-extrusion machined from 2024-T62 aluminium and whose purpose is to distribute flight loads from the fuselage upper skin to the centre fuselage structure – Figure D-4 (Figure 6-12 from Lockheed Martin Corporation (1997)). High positive g-loads cause fatigue cracking in the

tab radii of the longeron. Each aircraft has 16 longeron inspection locations (8 tab radii for the LH longeron and 8 tab radii for the RH longeron). The plate thickness of the longeron is 0.090 inch (2.3 mm). Part preparation consists of removing the access covers and removing loose paint and form-in-place gasket material (thickness about 2 to 3 mm) with a non-metallic scraper.

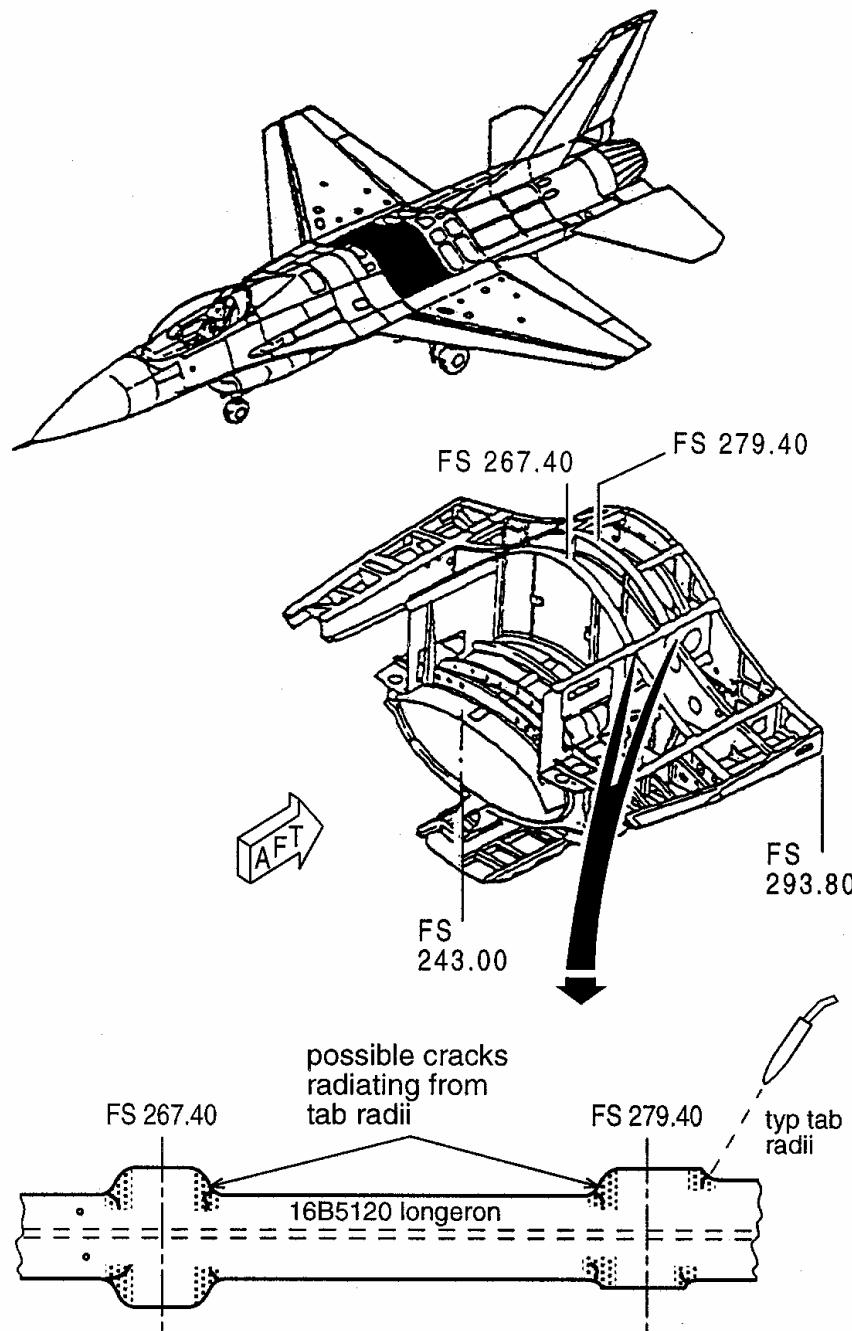


Figure D-4: Manual Eddy Current Inspection of the Tab Radii in the F-16 Centre Fuselage Longeron
(Figure 6-12 from Lockheed Martin Corporation (1997)).

b) Inspection Technique

NDI of the tab radii involves a manual eddy current inspection technique using standard phase-analysis equipment and standard eddy current probes (Figure D-4). In practice, air force bases apply different probes, but always at least two probes: 1) for crack detection *at* the radius, a stepped differential probe or a 45-degree shielded probe is used; and 2) for crack length measurement *along* the radius, a standard surface pencil-probe. The eddy current test frequency is 200 kHz. Calibration of inspection is done by adjusting the signal response of an EDM (electric discharge machined) surface notch with a depth of 0.020 inch (0.5 mm) to a 60% screen displacement of the phase-analysis instrument. The value for the reliably detectable crack size has been set at a through-crack with a length of 0.10 inch (2.5 mm).

Surface crack length is determined by applying parallel scan paths with the pencil-probe across a detected crack and marking the crack tips, and by measuring the crack length with a digital display calliper rule. The accuracy of this measurement is about 0.04 inch (1 mm) crack length.

c) CAMS Database

The inspection results of the F-16 longeron tab radii are stored in the CAMS (Core Automated Maintenance System) database. CAMS is an on-line and real-time system developed by the USAF to automate the most relevant aspects of the maintenance process. It is in use by the RNLAf for the registration of the status, utilisation, inventory, configuration and maintenance data of all RNLAf materiel (aircraft, engines, avionics, etc.). More specifically, CAMS is used as a comprehensive registration system for the ASIP (Aircraft Structural Integrity Program) field inspection data of F-16 aircraft – the longeron tab radii is one of the F-16 ASIP control points. When cracks are detected during the inspection of an ASIP point, then the number of cracks and the length of the largest crack found are registered in the CAMS database. The NDI signal responses are not recorded, so the NDI database is of the “hit” type.

The available CAMS field inspection data of the longeron tab radii (status May 2000) are shown in Tables D-2 to D-4. These tables list the actual crack lengths detected (values given in bold print) for 39 aircraft. It is noted, however, that these values suggest a high accuracy in crack length measurement (compare for example the values of 0.039 and 0.04 inch), which is not justified by the actual method of crack length measurement (accuracy about 0.04 inch). Therefore the reliability of the crack length values given in Tables D-2 to D-4 is lower than it might seem.

d) Crack Growth Data

For the longeron tab radii, a crack growth curve is available – Figure D-5 (from Lockheed (1993)). It is in fact a durability crack growth curve with an initial corner crack size of 0.007 x 0.007 inch (0.18 x 0.18 mm) and a functional impairment crack size of 0.187 inch (4.7 mm). The durability life represents the life during which flaws will not grow to an extent that requires extensive repair before one design service life. The longeron is treated as a durability item (and not as a damage tolerance item) because the longeron is believed not to be a safety-of-flight structure. The current inspection interval is 200 flight hours.

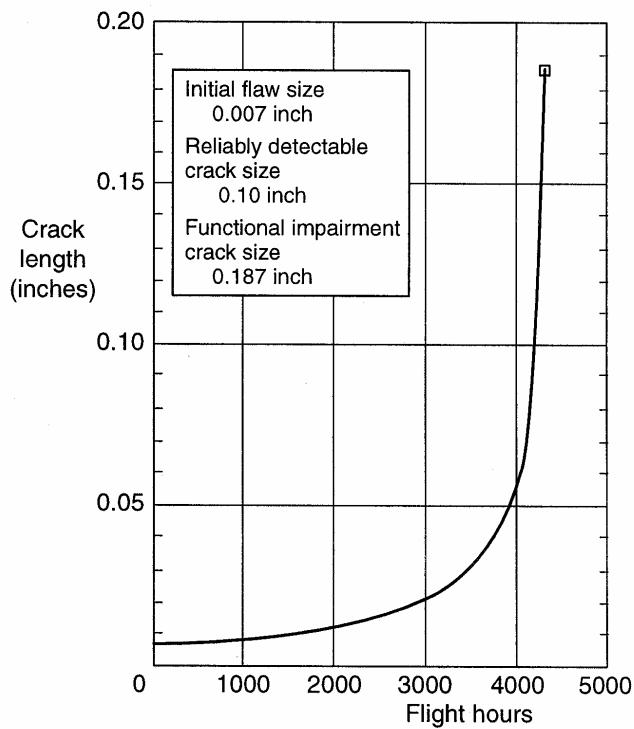


Figure D-5: Durability Crack Growth Curve for the F-16 Centre Fuselage Longeron Tab Radii (Figure 8.2.2-2 from Lockheed (1993)).

D.3.2 Inspection Databases

The durability crack growth curve of Figure D-5 has been used to estimate the previously missed crack sizes for each crack detected, using the back-extrapolation methodology discussed in Chapter 4.2. This results in an NDI database of the “hit/miss” type (Table D-2). This table lists the actual crack lengths detected and an estimation of the crack lengths missed (between brackets) for 39 aircraft. The CAMS database allowed the determination of 51 certain previous inspection times, resulting in a “hit/miss” data set of 90 data points.

Table D-2: Available CAMS Field Inspection Data of the Tab Radii of the F-16 Centre Fuselage Longeron (status May 2000). Listing of the actual crack lengths detected and estimation of crack lengths missed during previous phased inspections (between brackets). Back-extrapolation with the crack growth curve assumed valid for the baseline usage for F-16 aircraft in The Netherlands.

Database: 39 hits, 51 misses

Aircraft Number	Phased Inspection Times (Flight Hours)									
	??	1200	1400	1600	1800	2000	2200	2400	2600	2800
1		*	*	*	*	*		0.049	N.C.	
2		*	*	*	*	*		0.03	N.C.	
3		*	*	*	*	*		0.11	N.C.	
4		*	*	*		[0.030]	[0.038]	0.05		
5		*	*		0.049	N.C.				
6		*	*	*	*	*	[0.029]	[0.035]	0.047	N.C.
7		*	*	*	[0.025]	0.03	N.C.	N.C.		
8		*	*	*	*	*	[0.038]	0.05	N.C.	
9		[0.019]	[0.021]	[0.025]	0.03	N.C.				
10		*	*	*	[0.025]	0.03	N.C.			
11		*	*	[0.021]	[0.025]	0.03				
12		*	*	[0.019]	[0.021]	[0.025]	0.03			
13		*	*	*	[0.025]	0.03	N.C.			
14		*	*	[0.021]	[0.025]	0.03				
15		*	*	*	*	*	[0.030]	0.039		
16		*	*	*	*	*	[0.026]	[0.031]	0.04	N.C.
17		*	*	*	*	*	[0.019]	[0.021]	[0.025]	0.03
18		*	*	*	*	[0.031]	0.04	N.C.		
19		*	*	*	[0.025]	0.03	N.C.			
20		*	*	*	*	*	[0.064]	0.15		
21		*	*	[0.038]	0.05	N.C.	N.C.			
22		*	*	[0.025]	0.03					
23		*	*	[0.025]	0.03					
24		[0.021]	[0.025]	0.03	N.C.					
25		*	[0.021]	[0.025]	0.03	N.C.				
26		*	*	[0.065]	0.157					
27		*	*	*	*	[0.017]	0.019	N.C.		
28		*	*	*	*	*	*	[0.053]	[0.080]	0.236
29		*	*	*	*	[0.026]	[0.031]	[0.039]	[0.051]	0.07
30	0.06	*	*	*	*	*	*	N.C.	N.C.	
31	0.03	*	*	*	*	*	*	N.C.	N.C.	
32	0.05	*	*	*	*	*	*	N.C.	N.C.	
33		*	*	*	[0.023]	[0.026]	[0.031]	0.04		
34	0.11	*	*	N.C.	N.C.	N.C.	N.C.			
35	0.07	*	*	*	*	N.C.	N.C.	N.C.	N.C.	
36		*	*	*	*	[0.031]	[0.039]	[0.051]	0.07	
37		*	*	*	[0.019]	[0.021]	[0.025]	0.03		
38	0.03	*	*	*	N.C.	N.C.	N.C.			
39	0.15	*	*	*	N.C.					

?? – Inspection data unknown

* – No inspection data available

N.C. – Inspection performed, no crack detected

Besides Table D-2, two other databases have been constructed:

- Table D-3: CSI Corrected Data

Database with the “misses” back-extrapolated using aircraft individual crack growth curves based on the recorded specific spectrum crack severity index (SCSI). These curves have been derived from the crack growth curve assumed valid for the baseline usage (Figure D-5) by incorporating individual SCSI ratios

(ratio defined as the individual SCSI divided by the SCSI valid for the baseline usage). As for Table D-2, this database includes 90 “hit/miss” data points.

Table D-3: Available CAMS Field Inspection Data of the Tab Radii of the F-16 Centre Fuselage Longeron (status May 2000). Listing of the actual crack lengths detected and estimation of crack lengths missed during previous phased inspections (between brackets). Back-extrapolation with aircraft individual crack growth curves based on the specific spectrum severity index (SCSI).

Database: 39 hits, 51 misses

Aircraft Number	Phased Inspection Times (Flight Hours)										SCSI ratio
	??	1200	1400	1600	1800	2000	2200	2400	2600	2800	
1		*	*	*	*	*	0.049	N.C.			0.81
2		*	*	*	*	*	0.03	N.C.			0.87
3		*	*	*	*	*	0.11	N.C.			0.87
4		*	*	*	[0.035]	[0.041]	0.05				0.78
5		*	*	0.049	N.C.						0.78
6		*	*	*	*	[0.033]	[0.038]	0.047	N.C.		0.80
7		*	*	*	[0.027]	0.03	N.C.	N.C.			0.87
8		*	*	*	*	[0.040]	0.05	N.C.			0.81
9		[0.021]	[0.024]	[0.027]	0.03	N.C.					0.91
10		*	*	[0.027]	0.03	N.C.					0.90
11		*	*	[0.023]	[0.027]	0.03	N.C.				0.93
12		*	*	[0.021]	[0.024]	[0.027]	0.03	N.C.			0.89
13		*	*	*	[0.028]	0.03	N.C.				0.70
14		*	*	[0.025]	[0.027]	0.03	N.C.				0.73
15		*	*	*	*	[0.033]	0.039				0.84
16		*	*	*	*	[0.029]	[0.034]	0.04	N.C.		0.81
17		*	*	*	*	[0.022]	[0.025]	[0.027]	0.03		0.75
18		*	*	*	*	[0.034]	0.04	N.C.			0.87
19		*	*	*	[0.028]	0.03	N.C.				0.67
20		*	*	*	*	[0.076]	0.15				0.73
21		*	*	[0.039]	0.05	N.C.					0.96
22		*	*	[0.027]	0.03	N.C.					0.80
23		*	*	[0.027]	0.03	N.C.					0.83
24		*	[0.025]	[0.028]	0.03	N.C.					0.69
25		*	*	[0.024]	[0.027]	0.03	N.C.				0.80
26		*	*	[0.069]	0.157	N.C.					1.00
27		*	*	*	[0.017]	0.019	N.C.				1.06
28		*	*	*	*	*	[0.058]				0.87
29		*	*	*	*	[0.030]	[0.036]	[0.043]	[0.053]	0.236	0.79
30	0.06	*	*	*	*	*	N.C.	N.C.	N.C.		0.95
31	0.03	*	*	*	*	*	N.C.	N.C.	N.C.		0.76
32	0.05	*	*	*	*	*	N.C.	N.C.	N.C.		1.00
33		*	*	*	[0.026]	[0.029]	[0.034]	0.04			0.86
34	0.11	*	*	N.C.	N.C.	N.C.	N.C.				0.89
35	0.07	*	*	*	*	N.C.	N.C.	N.C.	N.C.	0.07	0.73
36		*	*	*	*	[0.038]	[0.045]	[0.054]	0.07		0.73
37		*	*	*	[0.022]	[0.025]	[0.027]	0.03			0.74
38	0.03	*	*	*	N.C.	N.C.	N.C.				0.77
39	0.15	*	*	*	N.C.						0.74

?? – Inspection data unknown * – No inspection data available N.C. – Inspection performed, no crack detected

- Table D-4: CSI Corrected Data, Extra Misses

Database with the “misses” also back-extrapolated using aircraft individual crack growth curves based on the recorded specific SCSI value. The difference with Table D-3 is that the 7 hits for which the inspection date

was unknown have been assigned an inspection date of 2800 flight hours, and that all 39 hits have been back-extrapolated to the arbitrary phased inspection time of 1200 flight hours. With this procedure, a total of 215 “miss” data points were determined, resulting in a “hit/miss” data set of 254 data points. The data set of 215 misses should be considered as an upper bound for the extent of the “miss” data set (based on the available number of hits). In reality, the size of the “miss” data set is smaller because the average first inspection time (for the longeron tab radii) is estimated at approximately 1400 to 1600 flight hours. The purpose of this analysis was to examine the influence of the “miss” data on the shape of the POD curves.

Table D-4: Available CAMS Field Inspection Data of the Tab Radii of the F-16 Centre Fuselage Longeron (status May 2000). Listing of the actual crack lengths detected and estimation of crack lengths missed during previous phased inspections (between brackets). Back-extrapolation with aircraft individual crack growth curves based on the specific spectrum severity index (SCSI). Back-extrapolation of all hits to the arbitrary phased inspection time of 1200 flight hours.

Database: 39 hits, 215 misses

Aircraft Number	Phased Inspection Times (Flight Hours)									SCSI ratio
	1200	1400	1600	1800	2000	2200	2400	2600	2800	
1	[0.023]	[0.026]	[0.029]	[0.034]	[0.040]	0.049	N.C.			0.81
2	[0.018]	[0.019]	[0.021]	[0.024]	[0.027]	0.03	N.C.			0.87
3	[0.028]	[0.033]	[0.040]	[0.050]	[0.065]	0.11	N.C.			0.87
4	[0.024]	[0.027]	[0.030]	[0.035]	[0.041]	0.05				0.78
5	[0.034]	[0.040]	0.049	N.C.						0.78
6	[0.019]	[0.020]	[0.023]	[0.026]	[0.028]	[0.033]	[0.038]	0.047	N.C.	0.80
7	[0.019]	[0.021]	[0.024]	[0.027]	0.03	N.C.	N.C.			0.87
8	[0.021]	[0.024]	[0.026]	[0.029]	[0.034]	[0.040]	0.05	N.C.		0.81
9	[0.019]	[0.021]	[0.024]	[0.027]	0.03	N.C.				0.91
10	[0.019]	[0.021]	[0.024]	[0.027]	0.03	N.C.				0.90
11	[0.019]	[0.020]	[0.023]	[0.027]	0.03					0.93
12	[0.018]	[0.019]	[0.021]	[0.024]	[0.027]	0.03	N.C.			0.89
13	[0.020]	[0.023]	[0.025]	[0.028]	0.03	N.C.				0.70
14	[0.020]	[0.022]	[0.025]	[0.027]	0.03					0.73
15	[0.019]	[0.020]	[0.023]	[0.025]	[0.028]	[0.033]	0.039			0.84
16	[0.018]	[0.019]	[0.021]	[0.023]	[0.026]	[0.029]	[0.034]	0.04	N.C.	0.81
17	[0.015]	[0.016]	[0.018]	[0.019]	[0.020]	[0.022]	[0.025]	[0.027]	0.03	0.75
18	[0.020]	[0.023]	[0.026]	[0.029]	[0.034]	0.04	N.C.			0.87
19	[0.021]	[0.023]	[0.025]	[0.028]	0.03	N.C.				0.67
20	[0.029]	[0.034]	[0.039]	[0.047]	[0.057]	[0.076]	0.15			0.73
21	[0.028]	[0.032]	[0.039]	0.05	N.C.	N.C.				0.96
22	[0.022]	[0.024]	[0.027]	0.03						0.80
23	[0.021]	[0.024]	[0.027]	0.03						0.83
24	[0.023]	[0.025]	[0.028]	0.03	N.C.					0.69
25	[0.020]	[0.022]	[0.024]	[0.027]	0.03	N.C.				0.80
26	[0.039]	[0.050]	[0.069]	0.157						1.00
27	[0.012]	[0.012]	[0.014]	[0.016]	[0.017]	0.019	N.C.			1.06
28	[0.021]	[0.024]	[0.027]	[0.031]	[0.037]	[0.046]	[0.058]	[0.082]	0.236	0.87
29	[0.020]	[0.022]	[0.025]	[0.027]	[0.030]	[0.036]	[0.043]	[0.053]	0.07	0.79
30	[0.017]	[0.019]	[0.020]	[0.023]	[0.027]	[0.030]	[0.037]	[0.046]	0.06	0.95
31	[0.015]	[0.016]	[0.018]	[0.019]	[0.020]	[0.022]	[0.025]	[0.027]	0.03	0.76
32	[0.016]	[0.017]	[0.019]	[0.020]	[0.024]	[0.027]	[0.032]	[0.039]	0.05	1.00
33	[0.019]	[0.020]	[0.023]	[0.026]	[0.029]	[0.034]	0.04			0.86
34	[0.020]	[0.022]	[0.025]	[0.028]	[0.033]	[0.039]	[0.049]	[0.065]	0.11	0.89
35	[0.021]	[0.023]	[0.026]	[0.028]	[0.032]	[0.038]	[0.045]	[0.054]	0.07	0.73
36	[0.023]	[0.026]	[0.028]	[0.032]	[0.038]	[0.045]	[0.054]	[0.057]	0.07	0.73
37	[0.018]	[0.019]	[0.020]	[0.022]	[0.025]	[0.027]	0.03			0.74
38	[0.015]	[0.016]	[0.018]	[0.019]	[0.020]	[0.022]	[0.025]	[0.027]	0.03	0.77
39	[0.024]	[0.026]	[0.029]	[0.033]	[0.039]	[0.047]	[0.057]	[0.076]	0.15	0.74

N.C. – Inspection performed, no crack detected

D.3.3 Influence of Spectrum Severity on Back-Extrapolated Crack Sizes

Comparison of Tables D-2 and D-3 shows that the values of the missed crack sizes (between brackets) do not differ that much, suggesting that incorporation of the spectrum severity in the crack growth curve does not have a great influence on the values of the back-extrapolated missed crack sizes. A further check was done by adaptation of the crack growth curve of Figure D-5 to SCSI values ranging from 0.80 to 1.30, and by back-extrapolating the missed crack sizes originating from a detected crack (“hit”) of size 0.15 inch. The results of these calculations are given in Table D-5.

Table D-5: Influence of Spectrum Severity (SCSI) on the Back-Extrapolated Crack Sizes from a Detected Crack of Size 0.15 inch, using a crack growth curve typical for the inspection of the tab radii of the F-16 centre fuselage longeron.

SCSI	Back-calculated crack sizes from a detected crack of size 0.15 inch							
	-7	-6	-5	-4	-3	-2	-1	0
0.80	0.028	0.031	0.036	0.041	0.049	0.060	0.078	0.15
0.85	0.027	0.030	0.034	0.040	0.048	0.058	0.076	0.15
0.90	0.026	0.029	0.033	0.038	0.046	0.056	0.075	0.15
0.95	0.025	0.028	0.031	0.037	0.045	0.055	0.074	0.15
1.00	0.024	0.027	0.030	0.036	0.043	0.054	0.073	0.15
1.05	0.023	0.026	0.029	0.035	0.042	0.053	0.071	0.15
1.10	0.022	0.025	0.028	0.033	0.040	0.051	0.070	0.15
1.15	0.021	0.024	0.028	0.032	0.039	0.050	0.069	0.15
1.20	0.020	0.023	0.027	0.031	0.038	0.049	0.068	0.15
1.25	0.020	0.022	0.026	0.030	0.037	0.048	0.067	0.15
1.30	0.019	0.021	0.025	0.029	0.036	0.047	0.066	0.15
Average (inch)	0.0232	0.0260	0.0297	0.0347	0.0421	0.0528	0.0715	0.15
Average (mm)	0.59	0.66	0.76	0.88	1.07	1.34	1.82	3.81
Std. Dev. (inch)	0.0031	0.0033	0.0035	0.0041	0.0045	0.0042	0.0039	0
Std. Dev. (mm)	0.08	0.08	0.09	0.10	0.11	0.11	0.10	0
COV	0.134	0.127	0.118	0.118	0.107	0.080	0.055	0

COV: Coefficient of Variation = (Standard Deviation) / (Average)

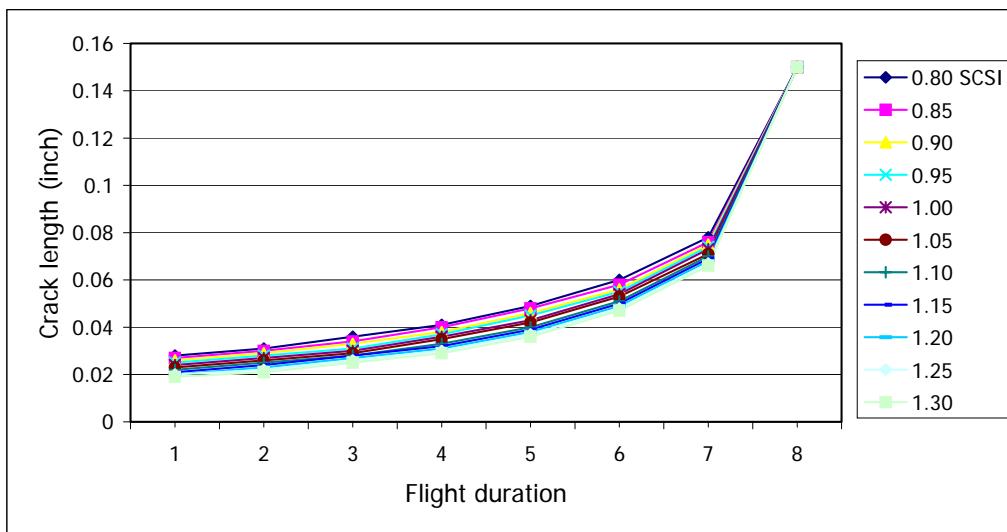


Table D-5 shows that the influence of spectrum severity on the back-extrapolated crack sizes is indeed relatively small. The standard deviation in these values, for an SCSI range of 0.80 – 1.30, first slightly increases to a maximum of 0.0045 inch (0.11 mm) and then decreases again. The coefficient of variation (COV), on the other hand, steadily increases for further back-extrapolating steps (mainly due to the lower absolute values of the back-extrapolated crack sizes).

These results suggest that the influence of spectrum severity on the CDF curve of back-extrapolated missed cracks is probably also small.

D.3.4 POD/CDF Curves for Different “Hit/Miss” Data Sets

POD and CDF curves have been drawn for six different “hit/miss” data sets using the data of Tables D-2 to D-4:

- a) Original data
Data of Table D-2: 39 hits, 51 misses
- b) SCSI corrected data
Data of Table D-3: 39 hits, 51 misses
- c) SCSI corrected data, extra misses
Data of Table D-4: 39 hits, 215 misses
- d) SCSI corrected data, without data of the largest crack (0.236 inch)
Data derived from Table D-3: 38 hits, 49 misses
- e) SCSI corrected data, without data of the smallest crack (0.019 inch)
Data derived from Table D-3: 38 hits, 50 misses
- f) SCSI corrected data, without data of the six cracks larger than 0.1 inch
(0.11, 0.11, 0.15, 0.15, 0.157 and 0.236 inch)
Data derived from Table D-3: 33 hits, 47 misses

The CDF-misses curves for the six different “miss” data sets, the CDF-hits curves for the four different “hit” data sets and the mean POD curves for the six different “hit/miss” data sets are given in Figures D-6 to D-8, respectively.

Figure D-6 shows that changes in spectrum severity, resulting in changes in the “miss” data set, have only a small influence on the CDF-misses curve (original data vs. SCSI corrected data), as was already indicated in Section D.2.3. A further observation is that changes in the “miss” data set, by adding or leaving out “miss” data, also have only a small influence on the CDF-misses curve.

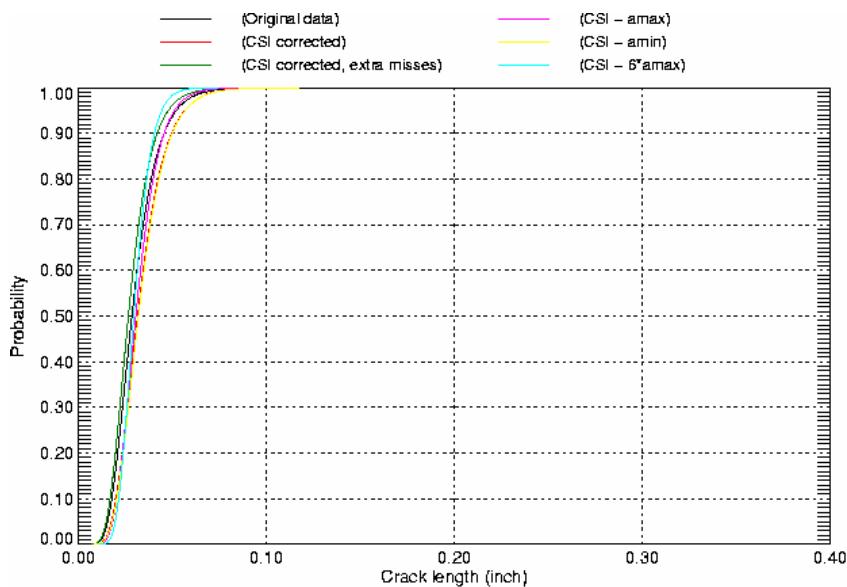


Figure D-6: CDF-Misses Curves for the Six Different ‘Miss’ Data Sets of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

Figure D-7 shows that small changes in the ‘hit’ data set have a medium influence on the CDF-hits curve. Only the curve drawn for the SCSI corrected data without the data of the six cracks larger than 0.1 inch (see curve (CSI – 6*amax)) has shifted significantly to the left of the curve valid for the original data. An arbitrary 90% probability criterion would yield the crack lengths of 0.059 and 0.107 inch for these curves, respectively. Leaving out a single data point has a small influence on the CDF-hits curve. For example, the data set without the largest crack (0.236 inch) results in a crack length of 0.094 inch for the 90% probability criterion (see curve (CSI – amax)).

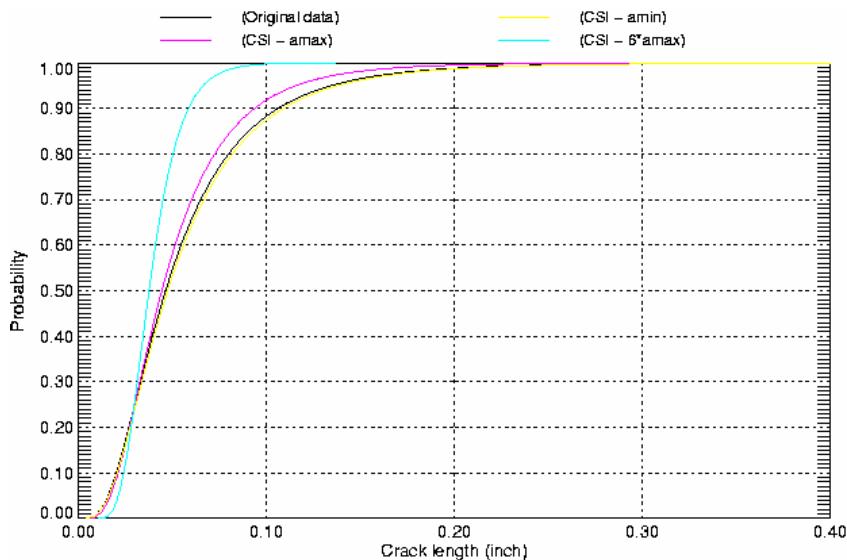


Figure D-7: CDF-Hits Curves for the Four Different ‘Hit’ Data Sets of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

Figure D-8 gives the *mean POD* curves for the six different “hit/miss” data sets. The figure shows that changes in spectrum severity (resulting in changes in the “miss” data set) and small changes in the “hit/miss” data set (by adding or leaving out “hit” or “miss” data) have a large influence on the mean POD curve. An arbitrary 90% probability criterion would yield crack lengths in the range of 0.069 inch (see curve (CSI – 6*amax)) to 0.140 inch (see curve (CSI corrected, extra misses)), when compared to the crack length of 0.089 inch for the original mean POD curve (data of Table D-2).

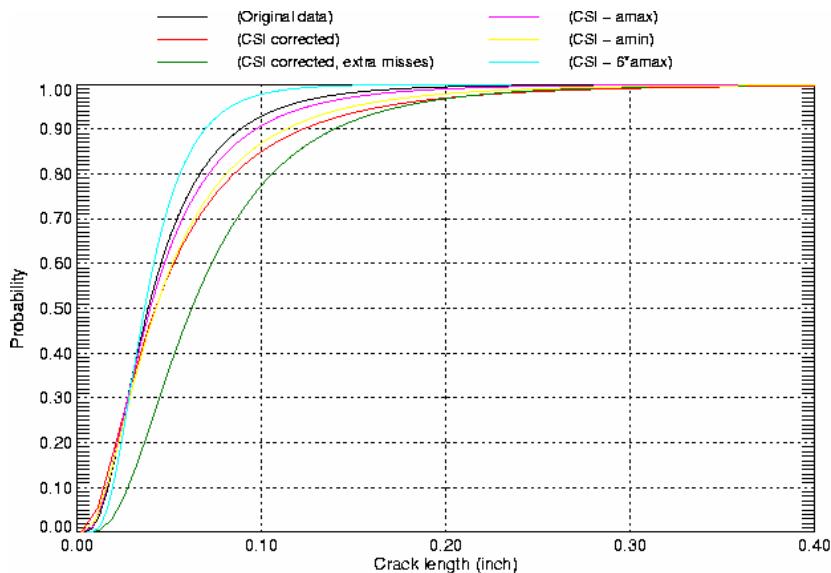


Figure D-8: Mean POD Curves for the Six Different “Hit/Miss” Data Sets of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

The preceding section shows that small changes in the “miss” or “hit” data set have only a small influence on the CDF-misses or CDF-hits curves, respectively. On the other hand, small changes in the “hit/miss” data set have a large influence on the mean POD curve. This implies that the production of a “reliable” POD curve will be very difficult in practice. This is because of the high unreliability in the values of the detected crack sizes and because of the unreliability in the values of the “miss” data (the generally used back-extrapolation procedure is unreliable, but has to be used owing to the unknown real crack growth curve). Thus, the CDF-hits curve is more stable and less vulnerable to changes in the data set than the POD curve.

The mean POD, CDF-hits and CDF-misses curves for the six different data sets are given in Figures D-9 to D-14. Table D-6 gives an overview of the relevant parameters of these curves, viz. the values of the mean (location parameter μ), standard deviation (scale parameter σ), a_{50} (crack length at 50% probability) and a_{90} (crack length at 90% probability). As can be expected, for the probability range of interest (probability larger than about 30%), all CDF-misses curves are located to the left of the mean POD curve. More importantly, however, the CDF-hits curves are not always located to the right of the mean POD curve, as was observed for the data of Fahr et al. (1995). Only the original “hit/miss” data set (Figure D-9) results in a CDF-hits curve located to the right of the mean POD curve, i.e. it is conservative here. In the other cases with differing “hit/miss” data, the CDF-hits curve is located close to the mean POD curve or located to the left of the mean POD curve, suggesting a non-conservative estimate of the detectable crack size. However, this is mainly due to the strong shift of the POD curve. The POD curve in Figure D-10, for example, shows a strong shift due to small changes in the values of the miss data (compare the locations of the x-symbols on the x-axis of

Figures D-9 and D-10), while the CDF-hits curve remains unchanged. The CDF-hits curve is much more stable than the POD curve and less vulnerable to changes in the data set.

Table D-6: Overview of the Relevant Parameters of the Mean POD, CDF-Hits and CDF-Misses Curves for the Six Different Data Sets given in Figures D-9 to D-14

Curve	Parameter [inch]			
	Mean	Std. Dev.	a50	a90
POD				
a	0.076	0.053	0.062	0.140
b	0.060	0.059	0.042	0.122
c	0.048	0.035	0.039	0.089
d	0.051	0.040	0.040	0.097
e	0.057	0.050	0.043	0.113
f	0.041	0.022	0.037	0.069
CDF-hits				
a	0.057	0.041	0.046	0.106
b	0.057	0.041	0.046	0.106
c	0.057	0.041	0.046	0.106
d	0.053	0.033	0.044	0.094
e	0.059	0.042	0.047	0.109
f	0.040	0.014	0.038	0.059
CDF-misses				
a	0.028	0.010	n.a.	n.a.
b	0.034	0.012	n.a.	n.a.
c	0.030	0.012	n.a.	n.a.
d	0.032	0.010	n.a.	n.a.
e	0.034	0.012	n.a.	n.a.
f	0.030	0.008	n.a.	n.a.

n.a. – Not applicable

Data sets:

- a Original data (Figure D-9)
- b SCSI corrected data (Figure D-10)
- c SCSI corrected data, extra misses (Figure D-11)
- d SCSI corrected data, without data of the largest crack (Figure D-12)
- e SCSI corrected data, without data of the smallest crack (Figure D-13)
- f SCSI corrected data, without data of the six cracks larger than 0.1 inch (Figure D-14)

Parameters:

- Mean Location parameter μ
- Std. Dev. Standard deviation; scale parameter σ
- $a50$ Crack length at 50% probability
- $a90$ Crack length at 90% probability

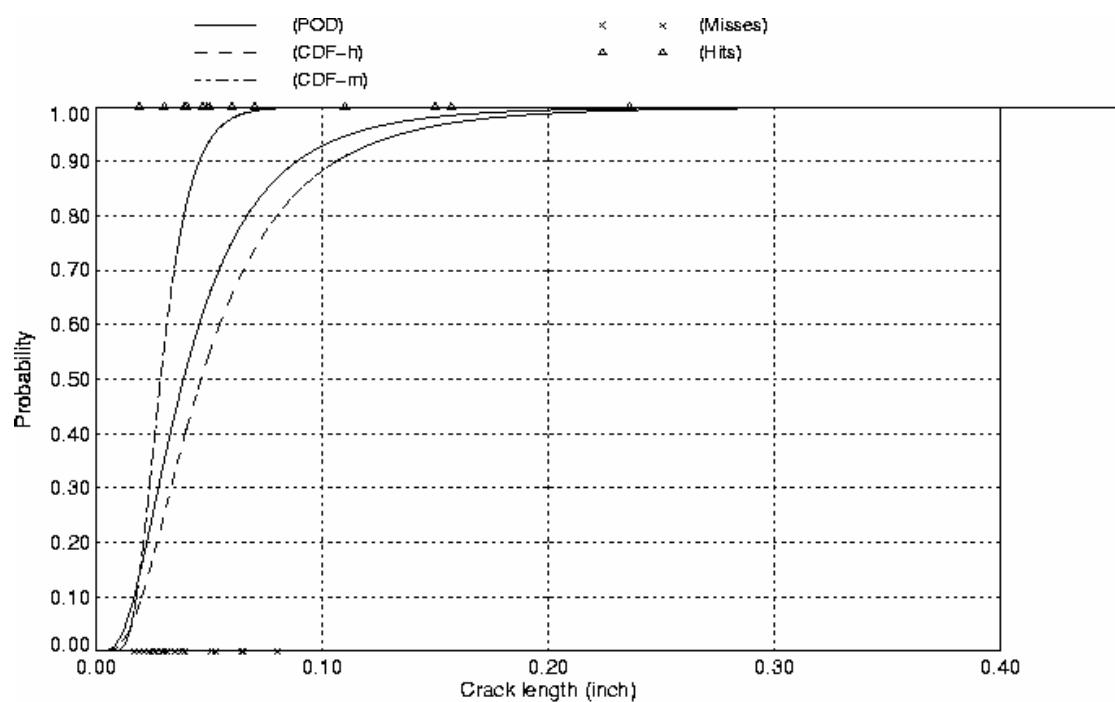


Figure D-9: Mean POD, CDF-Hits and CDF-Misses Curves for the Original “Hit/Miss” Data Set (39 hits, 51 misses) of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

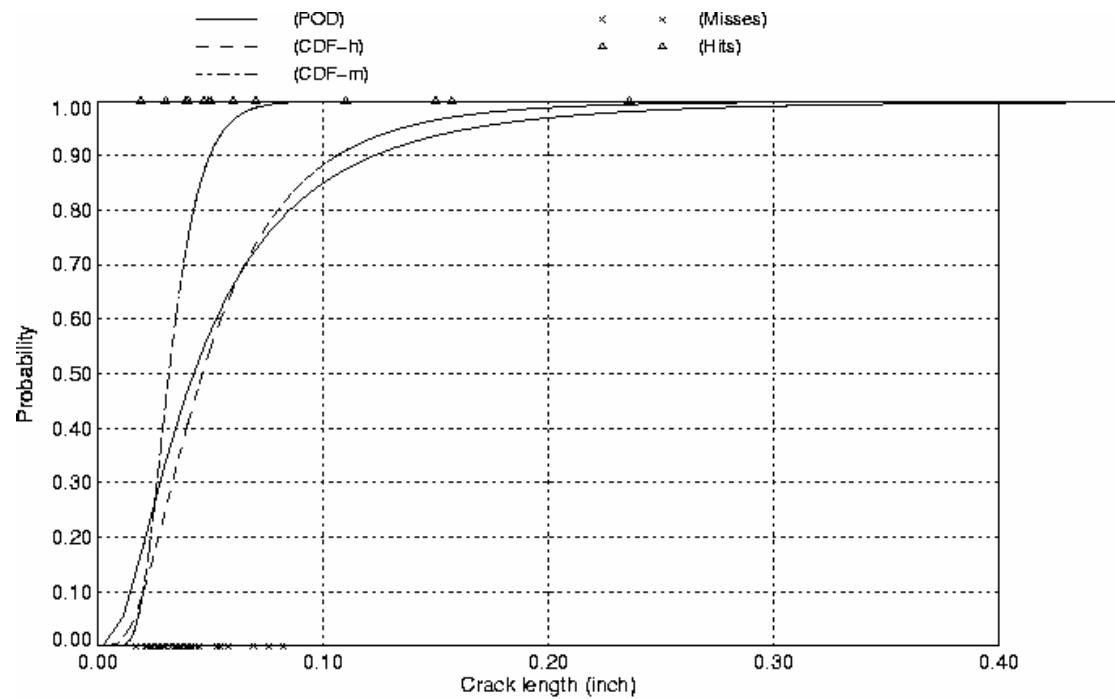


Figure D-10: Mean POD, CDF-Hits and CDF-Misses Curves for the SCSI Corrected “Hit/Miss” Data Set (39 hits, 51 misses) of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

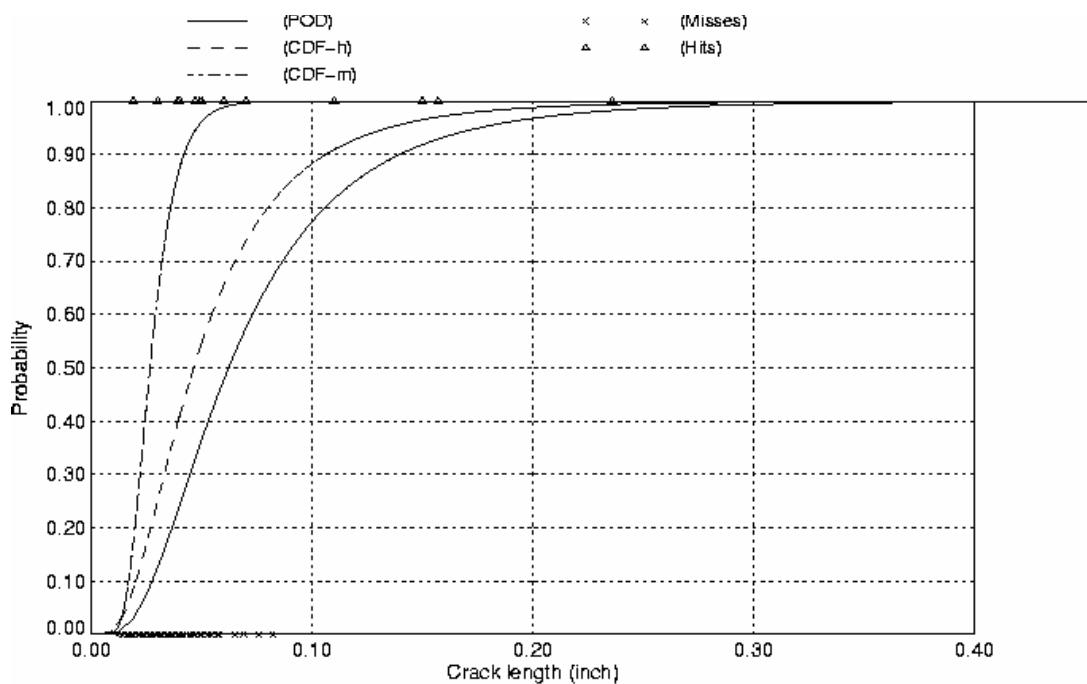


Figure D-11: Mean POD, CDF-Hits and CDF-Misses Curves for the SCSI Corrected “Hit/Miss” Data Set with Extra Misses (39 hits, 215 misses) of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

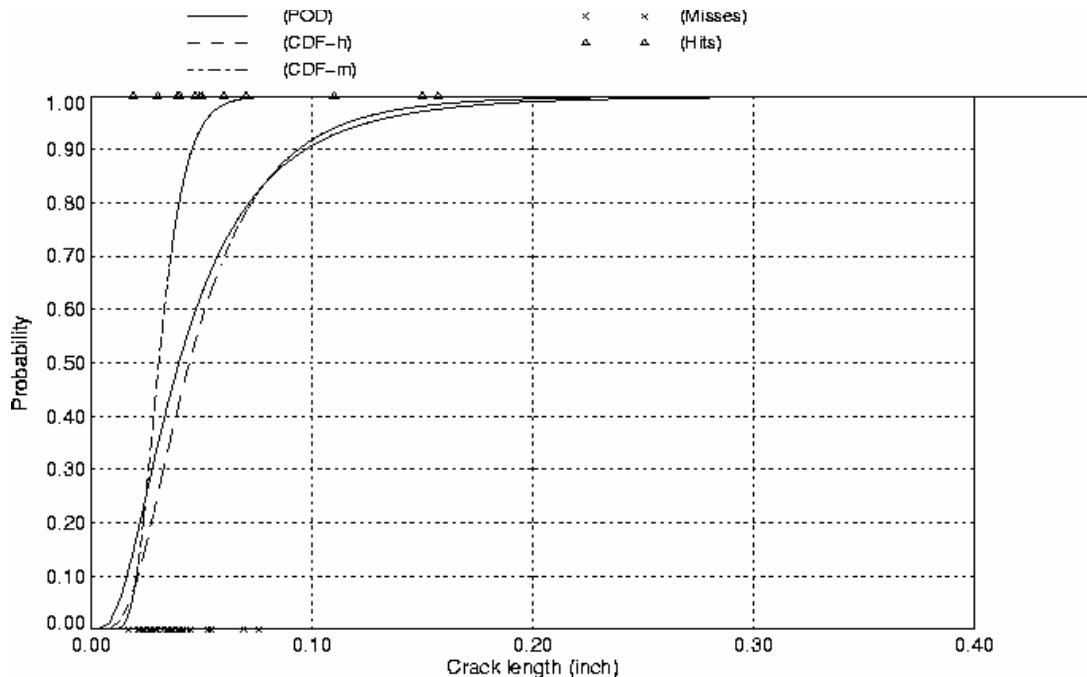


Figure D-12: Mean POD, CDF-Hits and CDF-Misses Curves for the SCSI Corrected “Hit/Miss” Data Set without Data of the Largest Crack (38 hits, 49 misses) of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

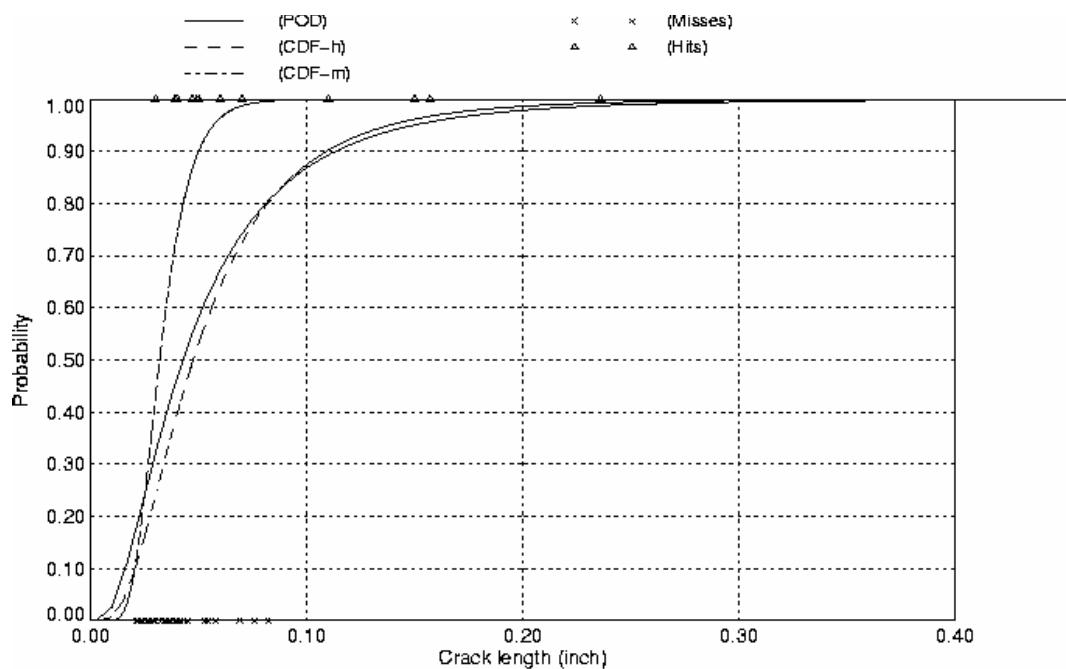


Figure D-13: Mean POD, CDF-Hits and CDF-Misses Curves for the SCSI Corrected “Hit/Miss” Data Set without Data of the Smallest Crack (38 hits, 50 misses) of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

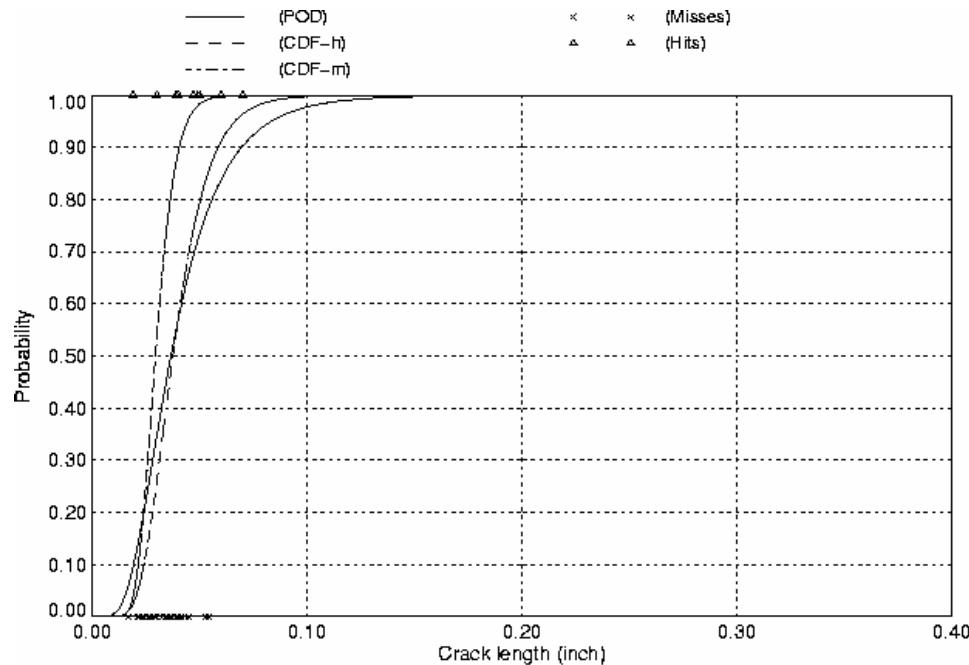


Figure D-14: Mean POD, CDF-Hits and CDF-Misses Curves for the SCSI Corrected “Hit/Miss” Data Set without Data of the Cracks Larger than 0.1 inch (33 hits, 47 misses) of the Manual Eddy Current Inspection of the F-16 Fuselage Longeron Tab Radii.

D.4 DISCUSSION

Information about the detectability of cracks in a field inspection environment can best be obtained with POD curves constructed from “hit/miss” data sets. However, it will be very difficult in practice to produce a “reliable” POD curve. This is caused by unreliability in the values of the detected crack sizes, by unreliability in the values of the “miss” data (back-extrapolation procedure in general) and because even small changes in the “hit/miss” data set can have a large influence on the POD curve. Further, for many inspection cases it will not be possible even to construct a “hit/miss” data set, for example in the absence of crack growth data, so that “miss” data points cannot be determined. In those cases, the CDF-hits curve can be of use. This curve is quite stable and less vulnerable to changes in the data set than the POD curve. It is emphasised that the CDF-hits curve is not a POD curve, but it does provide information about the detectability of cracks in a field inspection environment. Furthermore, it can give a first estimate of the detectable crack size.

D.5 CONCLUSION

The CDF-hits curve has a shape similar to the POD curve. It is **not** the POD curve, but it does provide information about the detectability of cracks in a field inspection environment. The CDF-hits curve does not directly yield the reliably detectable crack size (at a given confidence level), but it gives a first estimate of this size.

**ANNEX D – CUMULATIVE DISTRIBUTION
FUNCTION (CDF) OF DETECTED CRACKS**



Annex E – EVALUATION OF SAMPLE SIZE, CRACK SIZE AND MODEL IN THE POD CHARACTERIZATION OF NDI CAPABILITY

E.1 SIMULATED INSPECTIONS

Probability of detection, POD, is the conditional probability of detecting a crack given its size. The dependence of POD on crack size is expressed in functional form as $POD(a)$. In practice, the true $POD(a)$ for a defined inspection is never known exactly for the target crack sizes of the system. However, simulated inspections can be used to investigate the effects of data adequacy and estimation procedures on the sampling distributions of the estimates of the parameters of a $POD(a)$ function. The utility of the simulated inspections is enhanced by the knowledge of the “true” parameter values of the $POD(a)$ function. Accordingly, as part of this study, inspections were simulated to investigate the following effects:

- The effect of the number of the cracks (sample size) on the estimates of a_{90} and $a_{90/95}$, where a_{90} is the crack size for which there is 90% detectability, i.e. $POD(a_{90}) = 0.9$, and $a_{90/95}$ is the upper 95% confidence bound on the estimate of a_{90} .
- The effect of the sizes of the cracks in the analysis on the estimates of a_{90} and $a_{90/95}$.
- The effect of wrongly assuming a log-normal $POD(a)$ model when the true model is Weibull.
- The effect of sizes of the cracks being inspected on the cumulative distribution of detected cracks.

Simulating inspections is a simple process that was performed within a Microsoft Excel spreadsheet. The “crack” sizes and the $POD(a)$ function for an inspection set are defined. To simulate an inspection of a crack with size a_i , a uniform random number between zero and one is selected. If the number is less than $POD(a_i)$, the crack is considered detected, otherwise it is missed. The process is repeated by choosing a uniform random number for each crack in the defined set of sizes.

In the studies directed at sample sizes and crack sizes, it was assumed that the $POD(a)$ function is log-normal with $POD(100) = 0.5$ ($\mu = \ln 100$) and $\sigma = 0.5$. These parameters yield $a_{90} = 190$ mils. The a_{50} value of 100 mils is arbitrary and can be scaled to other median detectability sizes. The σ value of 0.5 is representative of a well-controlled, semi-automated, eddy current inspection. A Weibull cumulative distribution was used to investigate the effect of a wrong model being fit to inspection results. The shape and scale parameters of the Weibull model were selected so that a_{50} was either 50 or 100 mils and $a_{90} = 190$ mils.

The crack sizes used in the simulated inspections were selected to represent populations of small, medium, large and very large cracks when comparing crack size to the $POD(a)$ capability. Specifically, random samples of 100, 300 and 500 cracks were selected from log-normal distributions with $\sigma = 0.5$ and medians of 50, 100, 150 and 300 mils. Only one random sample of crack sizes was used for each combination of sample size and median crack size. Several preliminary simulation runs indicated that the effect of selecting new crack sizes for each simulated inspection was not significant.

E.2 CRACK SIZE AND SAMPLE SIZE EFFECTS ON ESTIMATES OF a_{90} AND $a_{90/95}$

In controlled NDI capability demonstrations, representative specimens with cracks of known sizes are inspected and the $POD(a)$ characterization is calculated from the inspection results. If most of the cracks in

the specimens are always found or always missed, very little information would be obtained and there would be a large degree of statistical uncertainty in the characterization of capability. The effective sample size for increasing the statistical validity of the characterization depends not only on the number of the cracks in the demonstration analysis, but also on their sizes. To the extent possible, the crack sizes in the specimens of a planned demonstration are selected to cover a target range of increase of the $POD(a)$ function.

In a service application, although an inspection system is selected to detect cracks in some target range of sizes, the sizes of the cracks that might be in the structure are independent of the inspection system capability. In the context of this study, there is no control over the sizes of the cracks in the structure being inspected, and thus, no control over the sizes of the cracks in the evaluation of the NDI system. To gain insight into the number of cracks that are needed to obtain reasonable precision in the characterization of capability, inspections were simulated for different combinations of crack size and sample size for an inspection with a known $POD(a)$ capability.

Ten sets of simulated inspections were generated in the sample size and crack size investigation. For each combination of crack sizes and sample size, 50 simulated inspections were generated using Microsoft Excel functions. The conditions for these simulations are defined in Table E-1.

Table E-1: Simulation Matrix for Crack Size and Sample Size Effects

# of Cracks	Crack Sizes			
	Small	Medium	Large	Very Large
100	50 Repeats	50 Repeats	50 Repeats	
300	50 Repeats	50 Repeats	50 Repeats	50 Repeats
500	50 Repeats	50 Repeats	50 Repeats	

POD: Log-normal – $\mu = \ln(100)$, $\sigma = 0.5$ ($a_{50} = 100$ mils, $a_{90} = 190$ mils)

Small Cracks: Random sample from log-normal – $a_{50} = 50$ mils, $\sigma = 0.5$

Medium Cracks: Random sample from log-normal – $a_{50} = 100$ mils, $\sigma = 0.5$

Large Cracks: Random sample from log-normal – $a_{50} = 150$ mils, $\sigma = 0.5$

Very Large Cracks: Random sample from log-normal – $a_{50} = 300$ mils, $\sigma = 0.5$

To show the location of the crack sizes with respect to the $POD(a)$ function, and to demonstrate the validity of the simulation process, the proportions of detected cracks in the 50 repeat runs of each crack size with the sample size of 300 were calculated and super-imposed on a plot of the assumed $POD(a)$ function. These comparisons are shown in Figure E-1 through to Figure E-4. All four figures demonstrate the agreement between the assumed $POD(a)$ function and the simulated detection proportions. Note in Figure E-1 that all but one of the small cracks were smaller than the a_{90} value of the $POD(a)$ function.

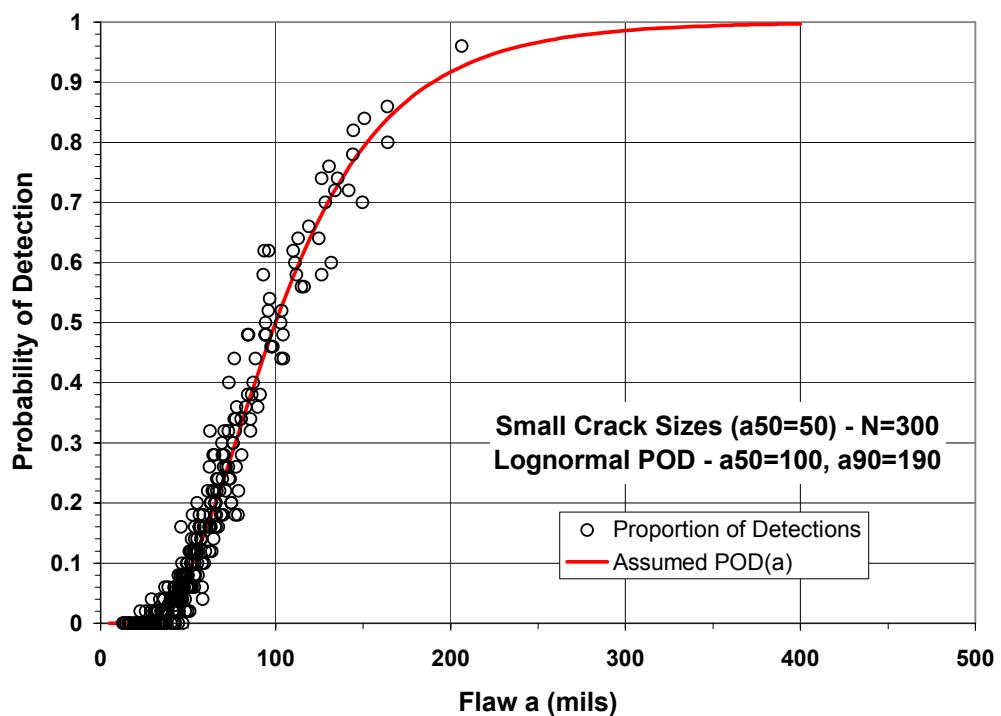


Figure E-1: Observed Proportion of Detections – Small Crack Sizes, $n = 300$.

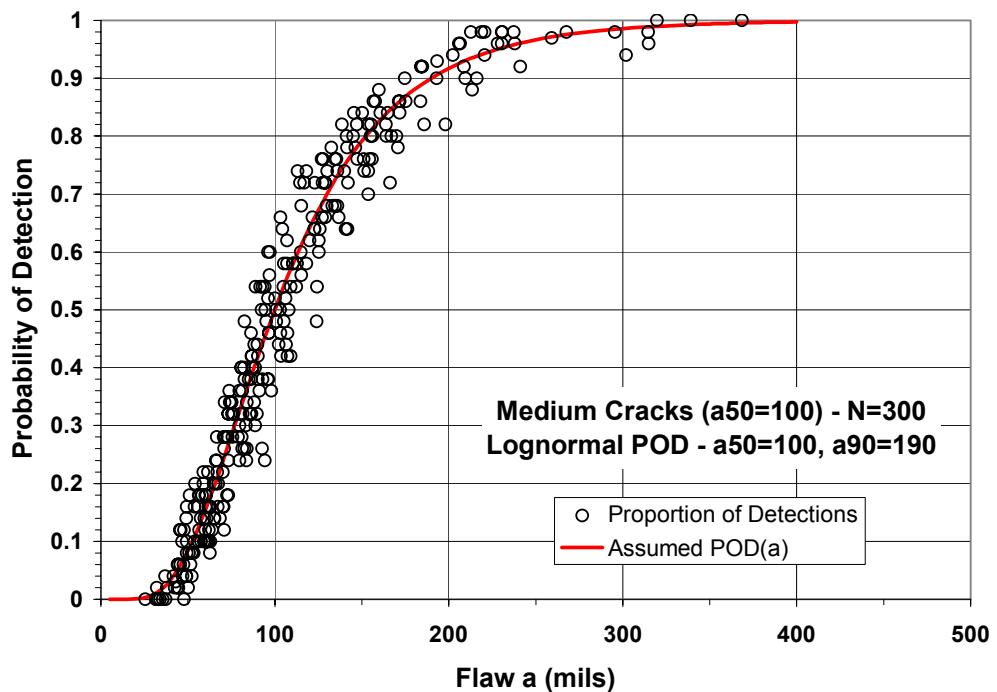


Figure E-2: Observed Proportion of Detections – Medium Crack Sizes, $n = 300$.

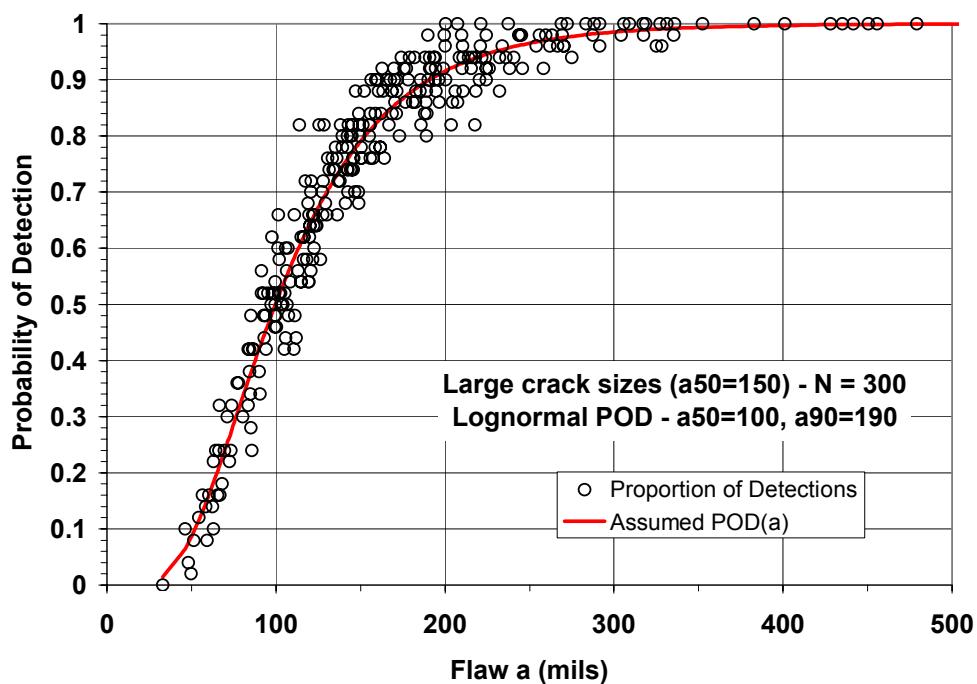


Figure E-3: Observed Proportion of Detections – Large Crack Sizes, $n = 300$.

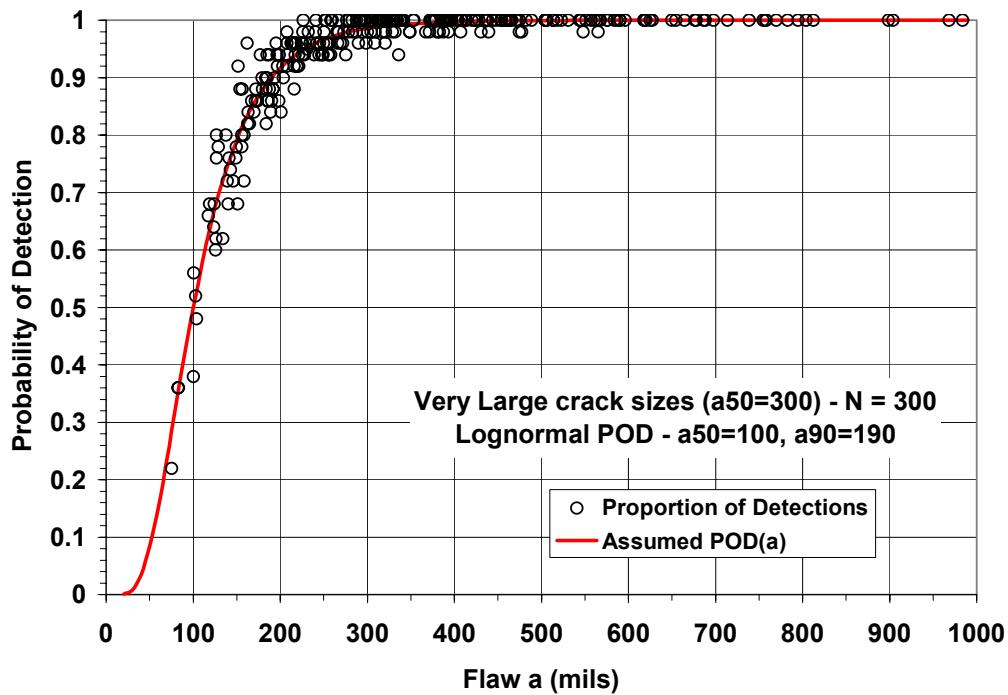


Figure E-4: Observed Proportion of Detections – Very Large Crack Sizes, $n = 300$.

For each of the 50 inspection sets of each combination of crack size and number of cracks, maximum likelihood estimates of the parameters of the cumulative log-normal $POD(a)$ function were calculated. The results were compared on the basis of the distributions of a_{90} and $a_{90/95}$ values for the combinations of crack size and sample size. Figure E-5, Figure E-6 and Figure E-7 compare the distributions of a_{90} and $a_{90/95}$ for the different crack sizes at sample sizes of 100, 300 and 500, respectively. To more easily evaluate the effects of sample size on a_{90} and $a_{90/95}$, the same distributions are rearranged and plotted in Figure E-8, Figure E-9 and Figure E-10 for the small, medium and large crack sizes, respectively.

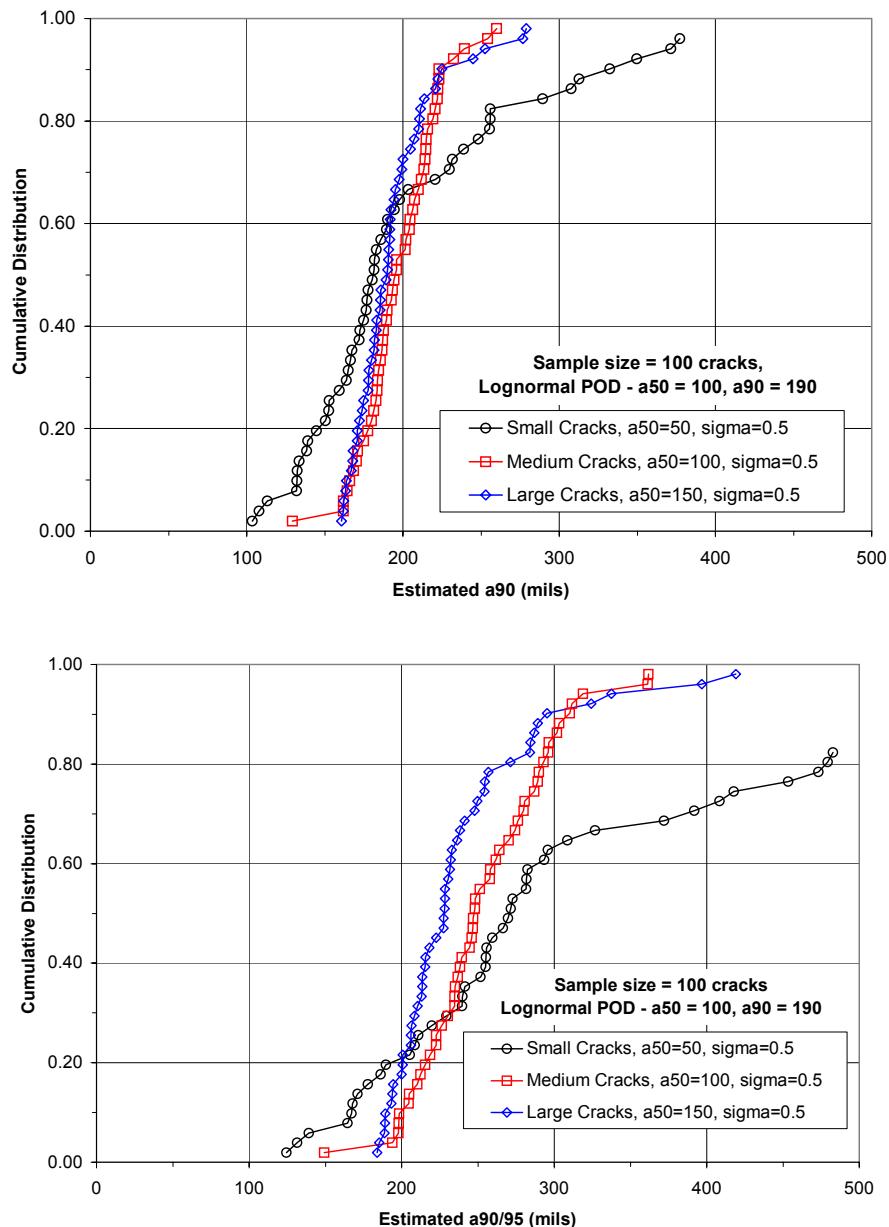


Figure E-5: Crack Size Effect for Samples Size of 100 Cracks:
Distributions of a_{90} (top) and $a_{90/95}$ for Small, Medium and Large Crack Specimens, $n = 100$.

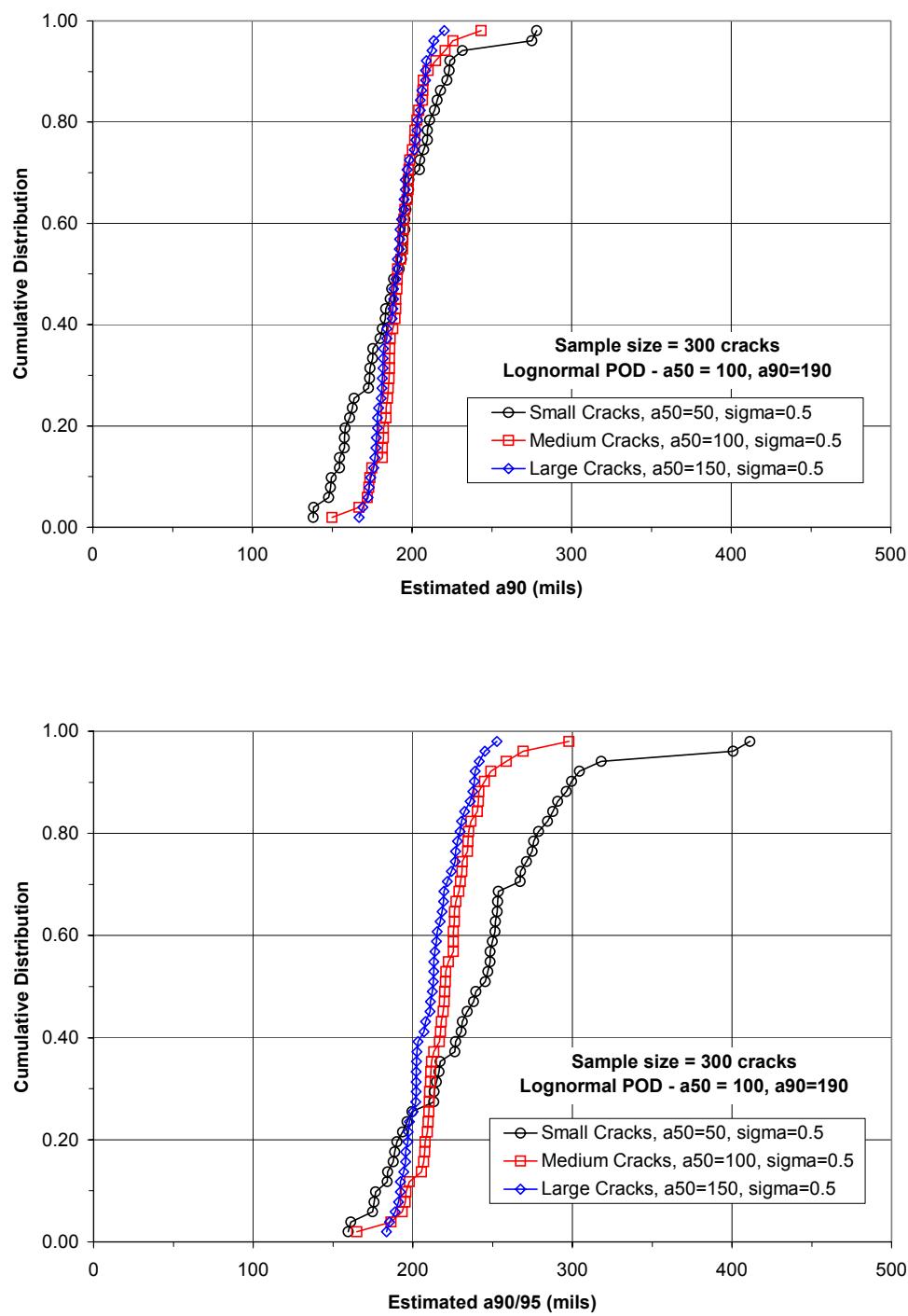


Figure E-6: Crack Size Effect for Samples Size of 300 Cracks:
Distributions of a_{90} (top) and $a_{90/95}$ for Small, Medium and Large Crack Specimens, $n = 300$.

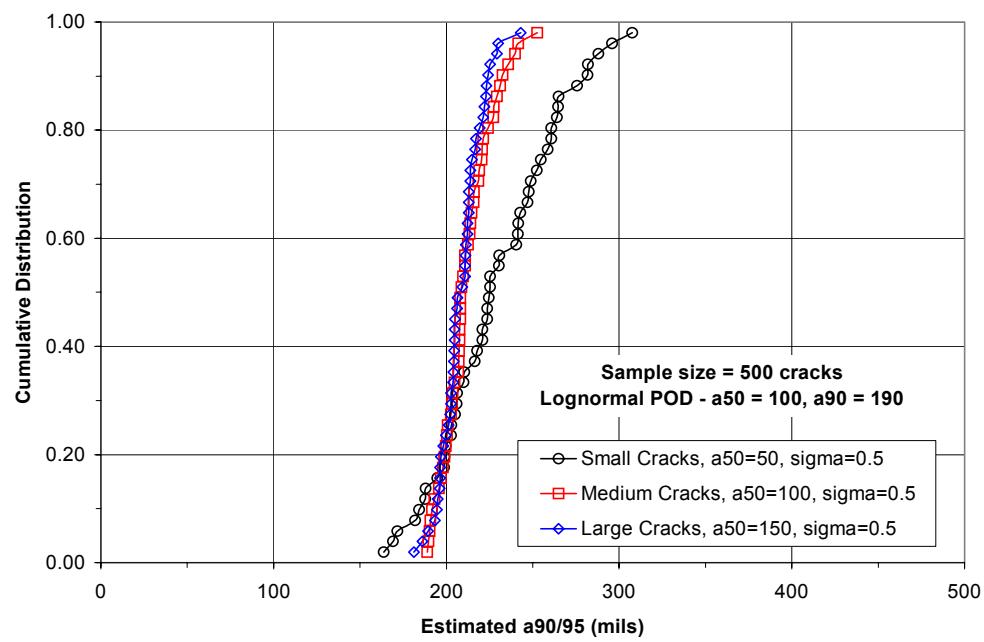
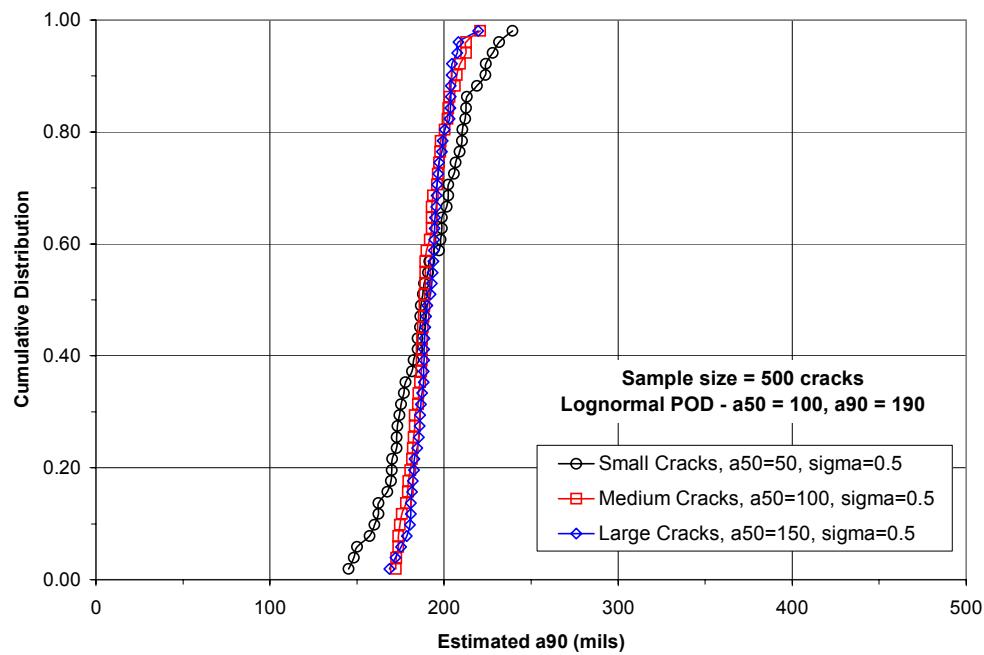


Figure E-7: Crack Size Effect for Samples Size of 500 Cracks:
Distributions of a_{90} (top) and $a_{90/95}$ for Small, Medium and Large Crack Specimens, $n = 500$.

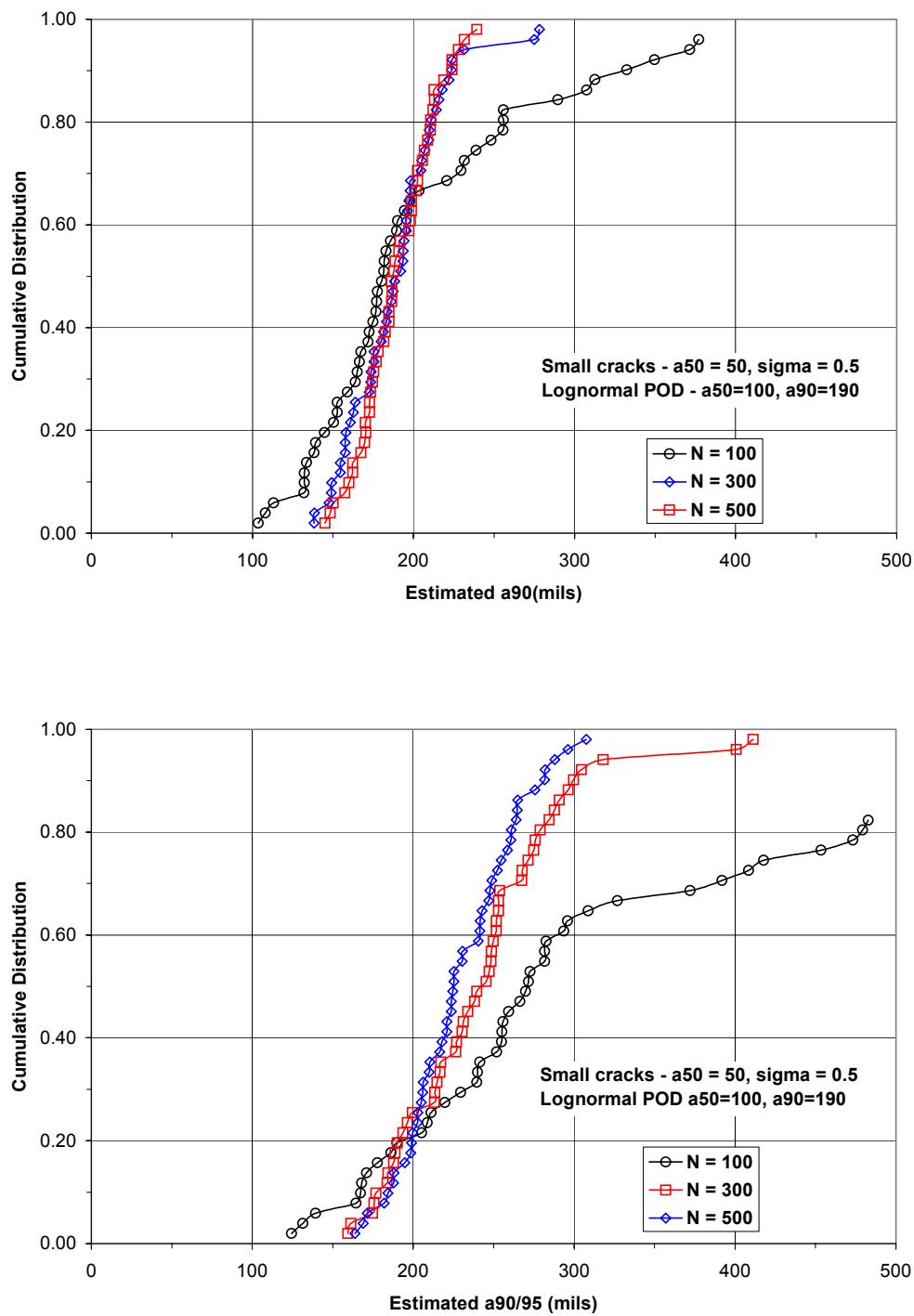


Figure E-8: Sample Size Effect for Small Crack Specimens:
Distributions of a_{90} (top) and $a_{90/95}$ for Small Crack Specimens, $n = 100, 300$ and 500 .

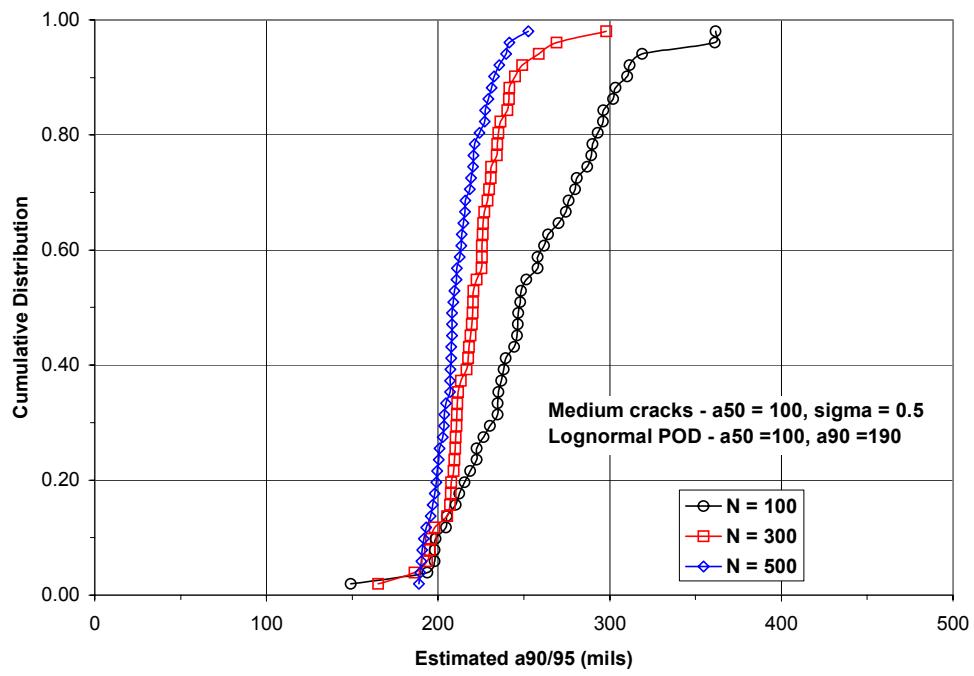
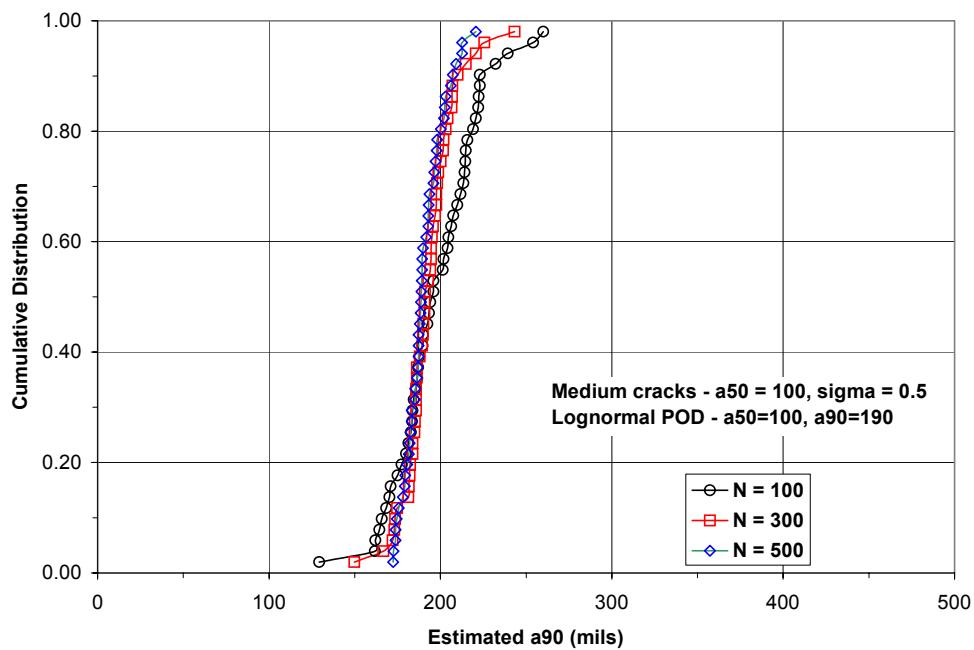


Figure E-9: Sample Size Effect for Medium Crack Specimens:
 Distributions of a_{90} (top) and $a_{90/95}$ Medium Crack Specimens, for $n = 100, 300$ and 500 .

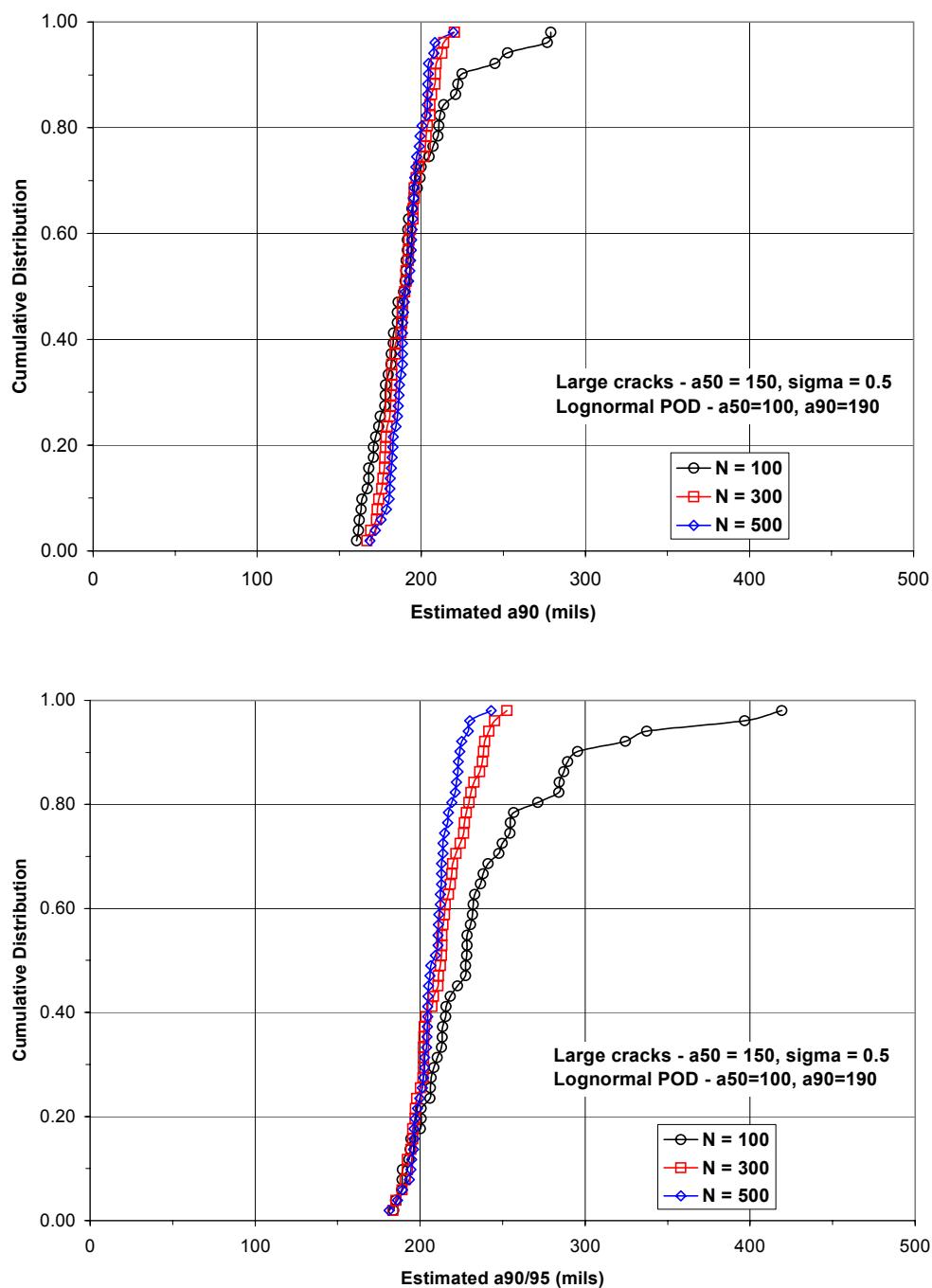


Figure E-10: Sample Size Effect for Large Crack Specimens: Distributions of a_{90} (top) and $a_{90/95}$ Large Crack Specimens, for $n = 100, 300$ and 500 .

The distributions of the a_{90} values are centered on the true a_{90} value of 190 mils. In eight of the nine simulations, the median a_{90} estimate is within 5 mils of the true value. In the simulation of inspections of 100 small cracks, the median a_{90} value was 10 mils (5%) less than the true value. It might be noted that the largest crack in the sample of 100 from the small crack distribution was 164 mils. The small crack effect of

very few inspections at or above the $POD(a)$ values of interest is also manifested in the increased scatter in the a_{90} estimates for the small crack inspections at all three sample sizes. The distributions of a_{90} for the medium and large crack sizes are equivalent for each of the three sample sizes.

The distributions of $a_{90/95}$ values also demonstrate the added information content when more of the inspected cracks are in the percentile of the $POD(a)$ function being estimated. There is significantly less scatter in the distributions of the confidence bounds for the medium and large crack sizes. The sample size effect is also apparent as the distributions of $a_{90/95}$ values are less conservative and display less scatter as sample size increases. For all three sample sizes, the percent of $a_{90/95}$ values less than the true value are reasonably close to 5% for the medium and large cracks. However, for the small cracks, the percent of $a_{90/95}$ values below 5% ranged were 20, 18 and 14 %, respectively, at $N = 100, 300$ and $N = 500$.

In a demonstration of inspection capability, the sizes of the cracks in the specimens to be inspected can be chosen by the evaluator. These simulations indicate the desirability of having the cracks centered on the a_{90} value if that is the parameter being used to characterize capability. However, in the analysis of data from in-service inspections, it would be expected that the unknown population of crack sizes would be small compared to the inspection capability. If not, a high percentage of the inspections would result in crack indications. Because the crack size population will be small when compared to the $POD(a)$ of the inspection system, a large sample size will be required to obtain stable estimates of either a_{90} or $a_{90/95}$. The results of Figure E-8 from the small crack simulated inspections indicate that 100 cracks of this relative size difference may not be sufficient to provide reasonable stability in the estimates of a_{90} or $a_{90/95}$. However, 300 cracks of this size would appear sufficient. Further simulations under more realistic conditions are warranted.

E.3 POD MODEL EFFECT ON ESTIMATES OF a_{90} AND $a_{90/95}$

A small simulation study was performed to consider the effect of fitting the wrong $POD(a)$ model when estimating the a_{90} crack size. In particular, inspection result data were generated assuming $POD(a)$ has the form of a Weibull cumulative distribution function, but the data were analyzed using a cumulative log-normal model. Two Weibull models were simulated. The parameters of the Weibull were determined so that $a_{90} = 190$ mils, with $a_{50} = 100$ and 50 mils. The first of these Weibull $POD(a)$ models, denoted as WBL-100, closely matches the log-normal of the crack size and sample size simulation study by having the same POD values at a_{50} and a_{90} . The second of the Weibull $POD(a)$ models, denoted as WBL-50, having a smaller a_{50} value at 50 mils, but the same a_{90} value at 190 mils, has a significantly different shape. The parameters of the $POD(a)$ models are given in Table E-2. All three $POD(a)$ models are shown in Figure E-11.

Table E-2: Parameter Values for $POD(a)$ Models of Simulation Study

Weibull $POD(a)$ Model

	a_{50}	a_{90}	Scale Parameter	Shape Parameter
WBL-100	100	190	121.6	1.87
WBL-50	50	190	75.2	0.90

Log-Normal $POD(a)$ Model

	a_{50}	a_{90}	Median	Standard Deviation
LN-100	100	190	ln(100)	0.5

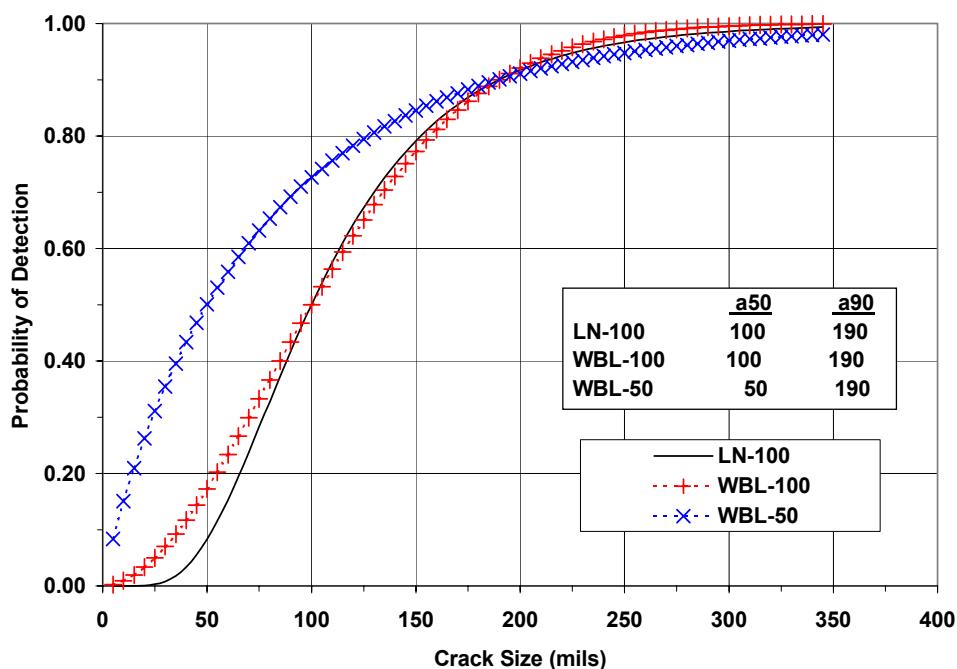


Figure E-11: POD(a) Models used in Simulated Inspections of Model Effect.

The sizes of the cracks in the simulated inspections from the Weibull POD(a) model were identical to those used in evaluation of the effects of crack size and sample size with the log-normal model. A sample size of 300 cracks was used in this evaluation of model effect. Given the Weibull models as “true,” fifty inspections were simulated for each of the 300 cracks from each of the three crack size distributions. The data were analyzed on the basis of the overall fit of the log-normal model to the “true” Weibull and the distributions of the a_{90} and $a_{90/95}$ values from the 50 simulated inspections.

Figure E-12 through to Figure E-16 present the proportion of detections of each crack for the two Weibull models and the three crack size distributions. Also shown on each plot are the “true” Weibull POD(a) function and the log-normal fit from a composite analysis of the 50 simulated inspections. In all six of the cases, the log-normal model agrees closely with the true Weibull model in the mid-ranges of the crack sizes in the analysis. However, the models disagree at the extremes of the data. In particular, the log-normal model produced a significantly larger estimate of a_{90} than the Weibull, when the sizes of the cracks in the simulated inspections are generally less than a_{90} (Figure E-12 and Figure E-13). The differences in a_{90} values between the models for the medium and large cracks are much less, but do reflect the lower upper tail values of the log-normal model. The log-normal model did reflect the change in shape of the two Weibull POD(a) functions.

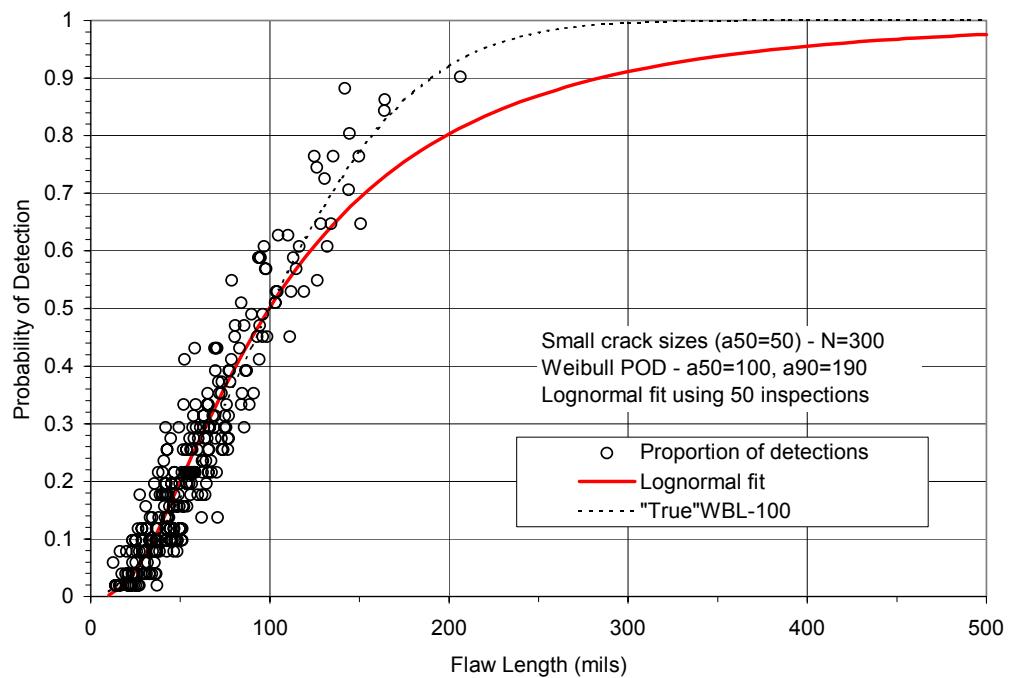


Figure E-12: Log-Normal Fit to WBL-100 POD – Small Crack Sizes, $n = 300$.

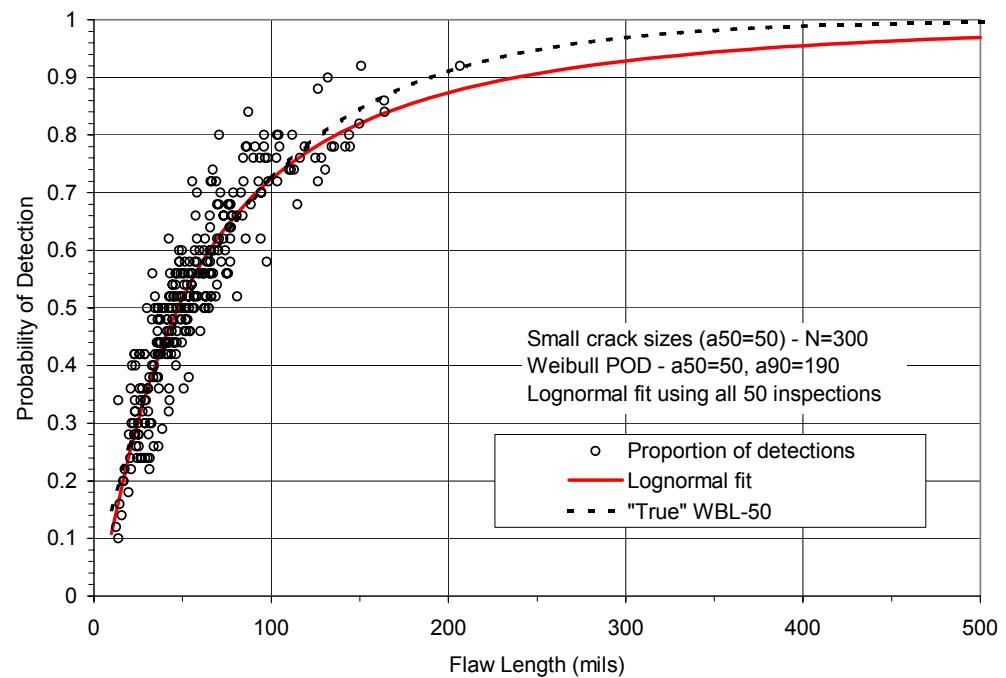


Figure E-13: Log-Normal Fit to WBL-50 POD – Small Crack Sizes, $n = 300$.

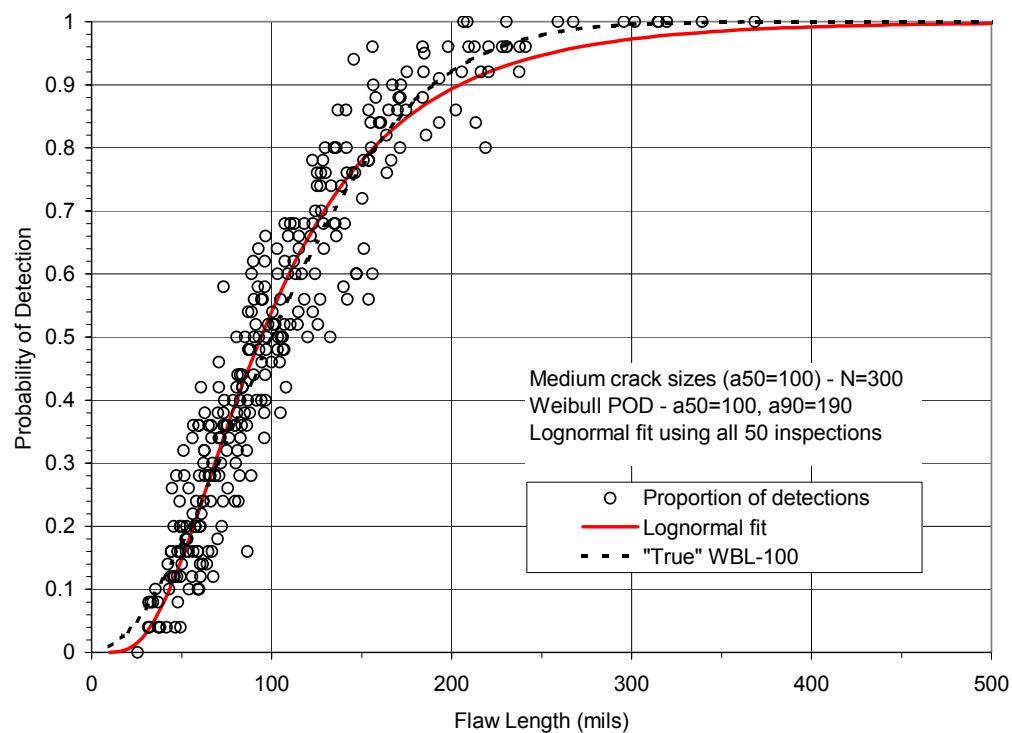


Figure E-14: Log-Normal Fit to WBL-100 POD – Medium Crack Sizes, $n = 300$.

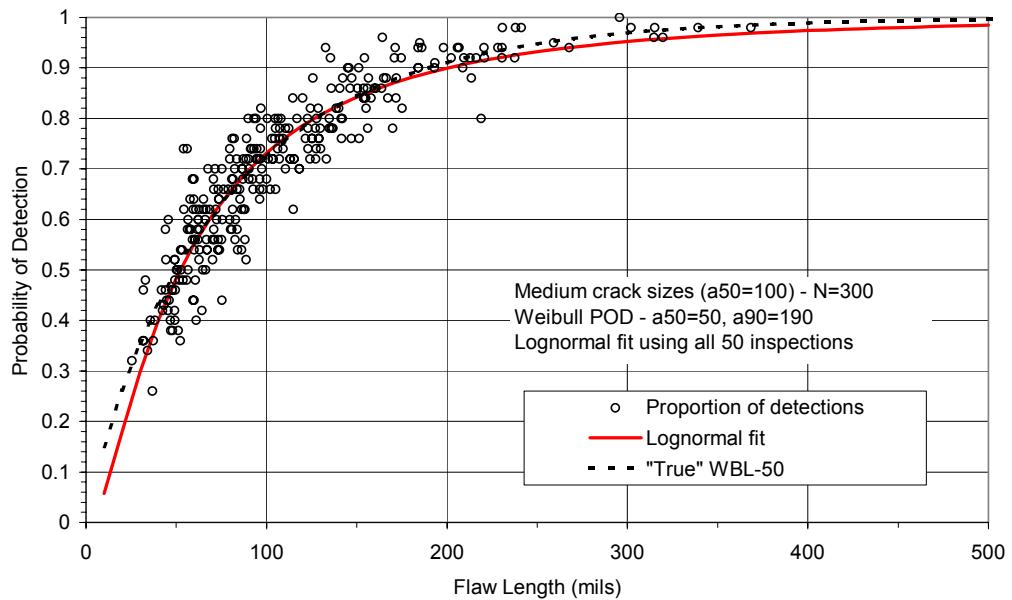


Figure E-15: Log-Normal Fit to WBL-50 POD – Medium Crack Sizes, $n = 300$.

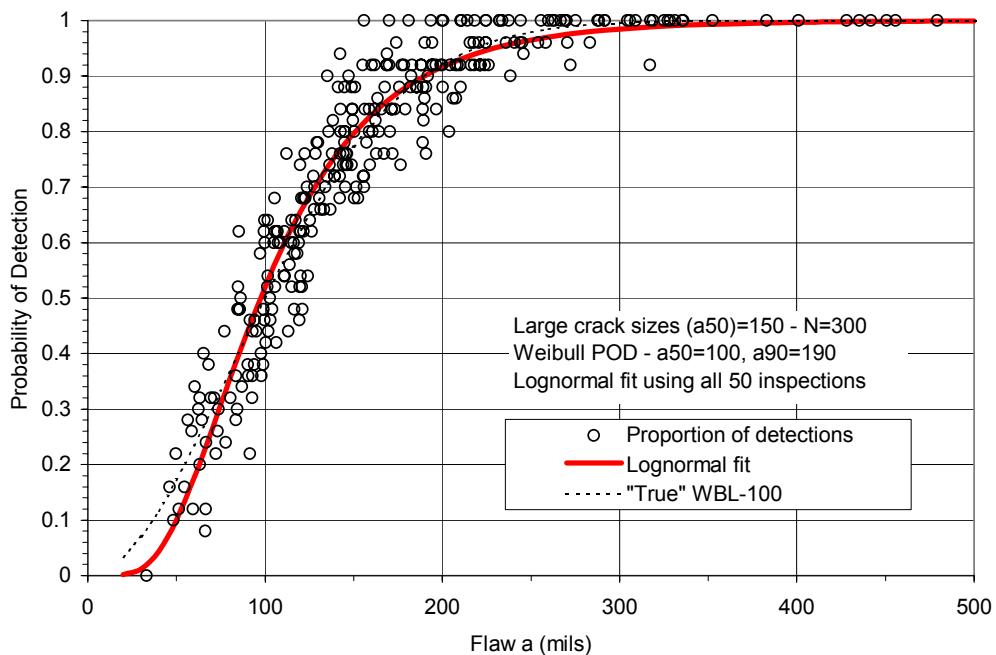


Figure E-16: Log-Normal Fit to WBL-100 POD – Large Crack Sizes, n = 300.

Distributions of a_{90} and $a_{90/95}$ from the log-normal fit to the Weibull $POD(a)$ functions are presented in Figure E-18 through to Figure E-20. Also included on the figures are the distributions of a_{90} and $a_{90/95}$ that were obtained when the $POD(a)$ was truly log-normal. In the simulated inspections of the small cracks, the log-normal fit to the Weibull $POD(a)$ yielded significantly larger (conservative) estimates of the true a_{90} value. The median estimate of a_{90} was 267 mils or 40% greater than the true value of 190 for WBL-100, the Weibull $POD(a)$ with $a_{50} = 100$. The median a_{90} was 224 mils or 18% greater than true for WBL-50, the Weibull $POD(a)$ with $a_{50} = 50$ mils. These results are consistent with the overall fits displayed in Figure E-12 and Figure E-13. Since the crack sizes are more in the increasing range of $POD(a)$ for WBL-50 than for WBL-100, there is less extrapolation in the estimate of a_{90} . The small crack $a_{90/95}$ values also display a large, significant model effect in the conservative direction.

Figure E-19 and Figure E-20 show that the model effect is lessened when the cracks in the inspections are closer to the a_{90} value. Figure E-20 shows that there is an insignificant model effect when the sizes of the cracks in the analysis cover the range of increase of the $POD(a)$ function and there are a large number of inspection results for cracks greater than a_{90} . The WBL-50 inspection simulations for the large cracks display more scatter in the estimates of a_{90} and $a_{90/95}$. This is likely due to an insufficient number of small cracks to define the $POD(a)$ shape in the small crack range (Figure E-17).

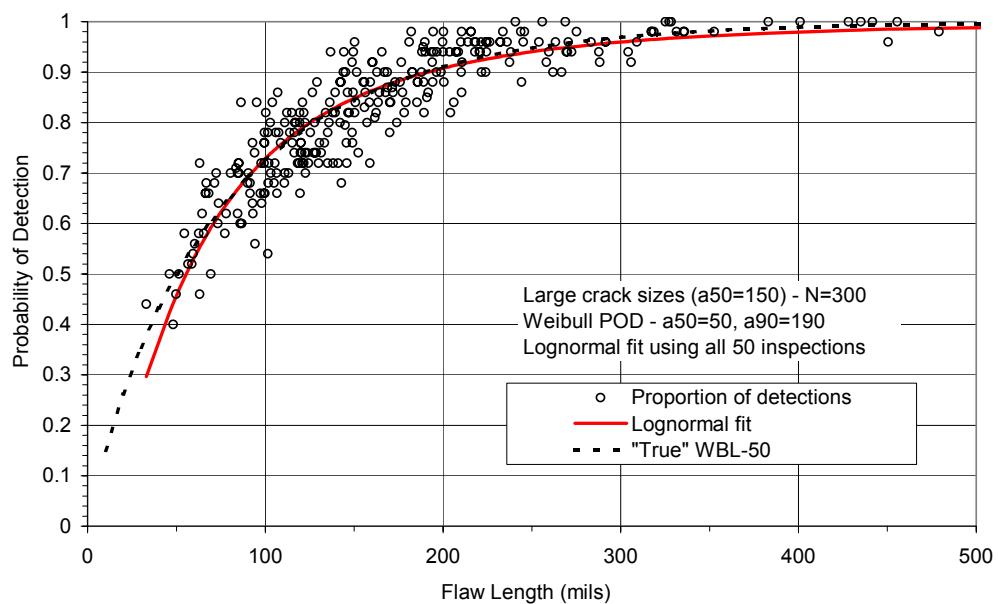


Figure E-17: Log-Normal Fit to WBL-50 POD – Large Crack Sizes, $n = 300$.

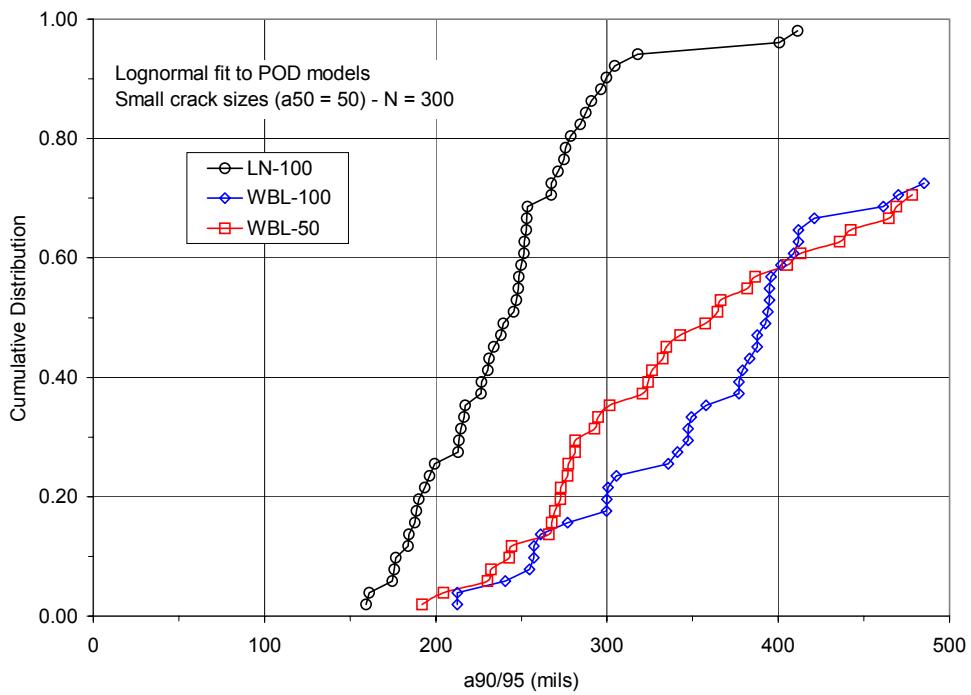
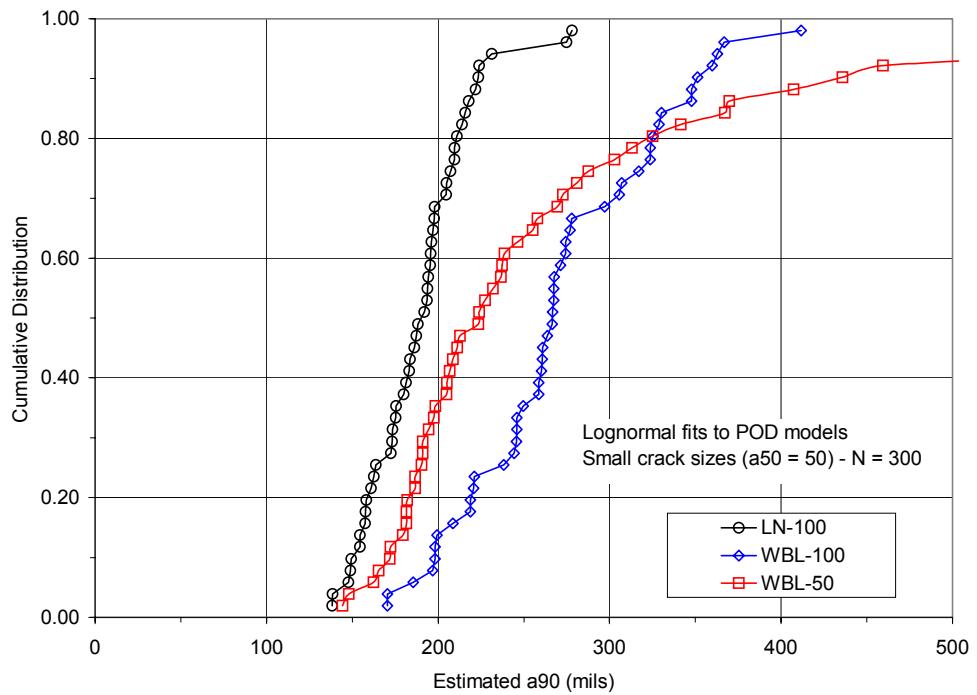
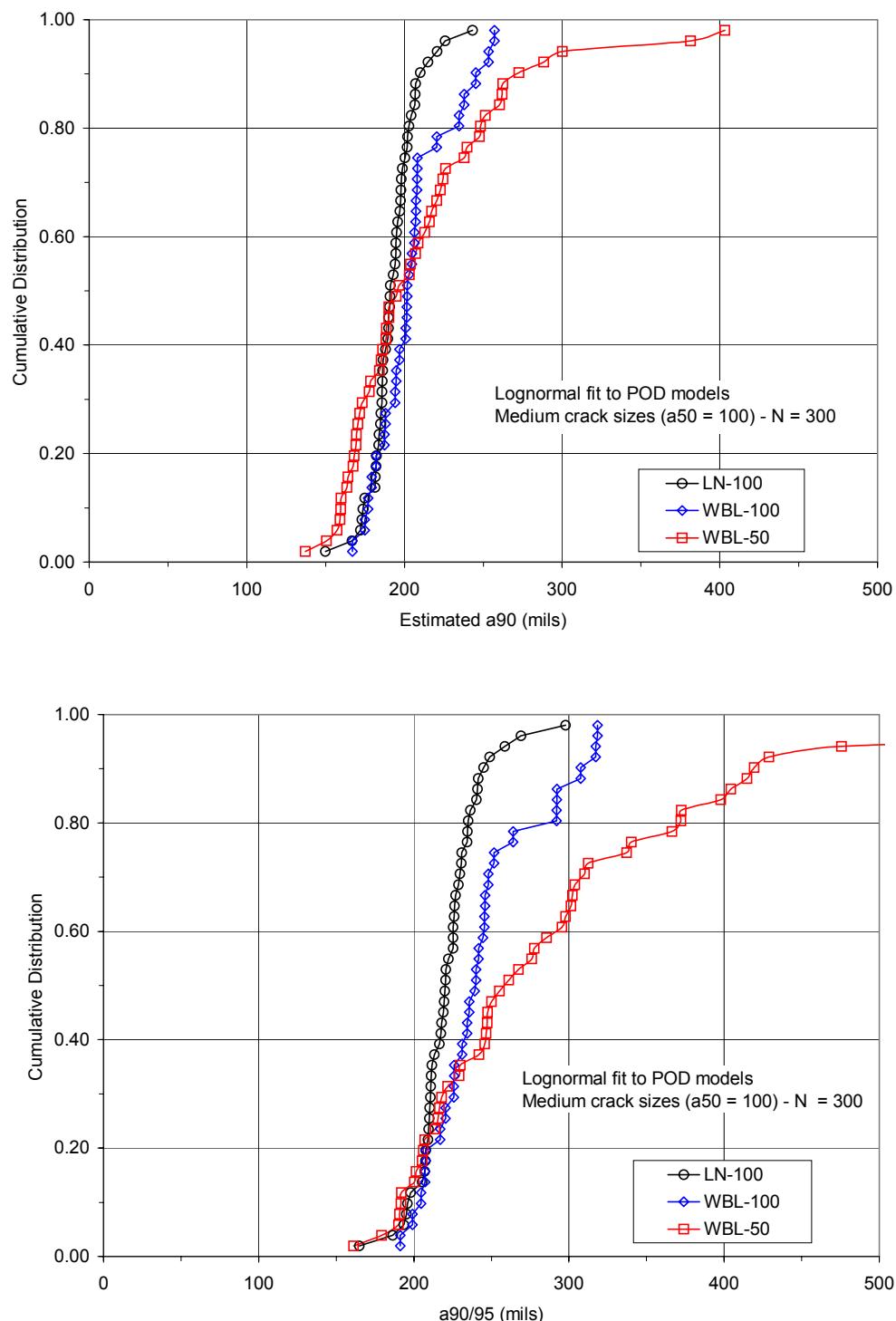
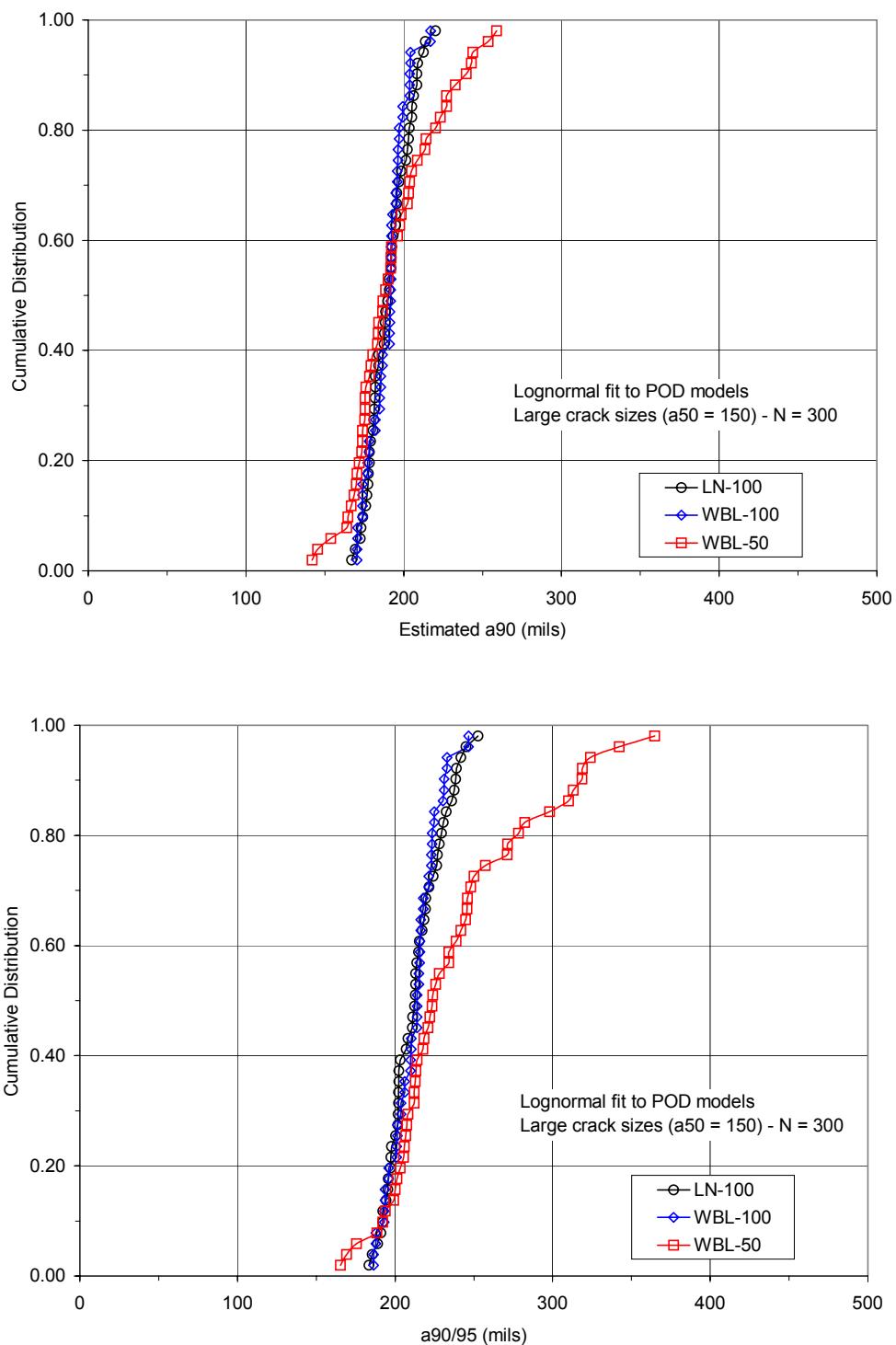


Figure E-18: POD Model Effect for Small Crack Specimens:
Distributions of a_{90} (top) and $a_{90/95}$ for Log-Normal Fit to POD Models.



**Figure E-19: POD Model Effect for Medium Crack Specimens:
Distributions of a_{90} (top) and $a_{90/95}$ for Log-Normal Fit to POD Models.**



**Figure E-20: POD Model Effect for Large Crack Specimens:
Distributions of a₉₀ (top) and a_{90/95} for Log-Normal Fit to POD Models.**

ANNEX E – EVALUATION OF SAMPLE SIZE, CRACK SIZE AND MODEL IN THE POD CHARACTERIZATION OF NDI CAPABILITY

The results of these simulations indicate that estimates of a_{90} are sensitive to the POD(a) model when there are few cracks of size a_{90} or greater. With mostly small cracks in the analysis, a_{90} is an extrapolation and sampling errors in parameter estimates are magnified. As discussed in the previous section, the crack sizes in the in-service inspections are expected to be small in comparison to a_{90} . If this assumption is true, very large sample sizes may be required to obtain estimates of a_{90} with reasonable precision.

Annex F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS

F.1 INTRODUCTION

The cracks that will be detected at in-service inspections depend on both the sizes of the cracks in the inspected structures as well as the efficacy of the inspection system. In general, neither of these is known. Because of the importance of the crack sizes in estimating the parameters of a $POD(a)$ function, a theoretical study was performed to investigate the effect of representative $POD(a)$ capabilities and pre-inspection crack size distributions on the distribution of the sizes of the detected cracks.

A distribution of crack sizes at a defined location is often used to represent the distribution of damage across a fleet. The distribution is defined in terms of a family of distributions, such as the log-normal family, whose parameters depend on the fatigue experience of the fleet. For the purposes of this study, it is assumed that the population of inspected cracks is log-normal and the parameters will be varied to reflect different sizes in relation to a $POD(a)$ capability.

The theoretical calculations for the distribution of the crack sizes detected at an inspection are as follows. Assume:

- $f(x)$ is the probability density function of crack sizes in the structure immediately before the inspection. $F(x)$ is the cumulative distribution function.
- $POD(x)$ is the probability of detecting a crack of size x .
- $G(a)$ is the proportion of cracks smaller than a that are detected.

Then:

$$G(a) = \int_0^a POD(x) f(x) dx. \quad (F-1)$$

$H(a)$ is the proportion of cracks smaller than a that are missed.

$$H(a) = \int_0^a [1 - POD(x)] f(x) dx. \quad (F-2)$$

$G(a) + H(a) = F(a)$, the proportion of all cracks smaller than a . $G(\infty)$ is the total proportion of inspections that result in a detection. Thus, the cumulative distribution of the sizes of the cracks detected during the inspection is given by the expression:

$$G_{det}(a) = G(a) / G(\infty). \quad (F-3)$$

The cumulative distribution of the sizes of the cracks that were missed during the inspection is given by the expression:

$$H_{miss}(a) = H(a) / [1 - G(\infty)]. \quad (F-4)$$

ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS

For this study, it is assumed that $POD(a)$ is log-normal, with parameters $\mu = \ln(a_{50})$ and σ . In the study, the 50% detectable crack size, a_{50} , is held constant at 50 mils (1.25 mm). The parameter σ is assigned values of 0.25, 0.5, 0.75, 1.0 and 1.25. These values are reasonably representative of semi-automated and manual eddy current inspections (NTIAC: DB-97-02, Non-destructive Evaluation (NDE) capabilities Data Book, Third Edition, Non-destructive Testing Information Analysis Center (NTIAC), Texas Research Institute Austin, Inc., November 1997). The five $POD(a)$ functions are shown in Figure F-1. The 90% detectable crack size, a_{90} , for these $POD(a)$ functions are 69, 95, 131, 180 and 248 mils for the five increasing σ values.

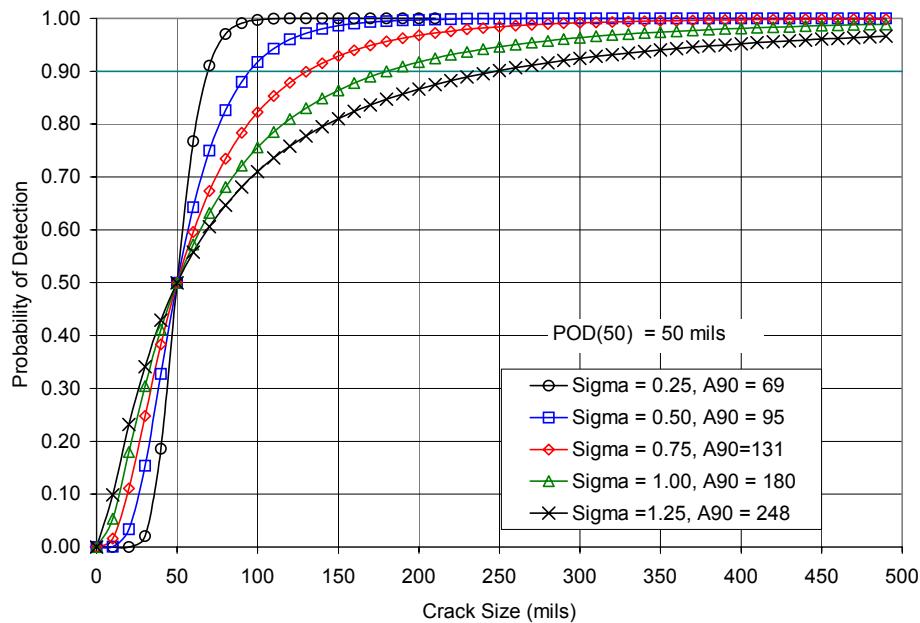


Figure F-1: POD Functions for Different σ values, Constant μ values.

It is assumed that the population of crack sizes in the structure also has a log-normal distribution. The standard deviation of the natural logarithm of crack sizes, σ , is assumed to be 0.75, 1.00 and 1.50. This degree of scatter has been used in structural risk analyses of military aircraft. Median crack sizes, a_{50} , were set arbitrarily at 5, 10, 20 and 30 mils. The probability density functions for the assumed crack size distributions with $\sigma = 0.75$ and median crack sizes of 10, 20 and 30 mils are shown in Figure F-2. As a size reference, the $POD(a)$ function with $a_{50} = 50$ and $\sigma = 0.5$ ($a_{90} = 95$ mils) is also included in the figure. Figure F-3 and Figure F-4 provide another view of the crack size distributions of this sensitivity analysis. Figure F-3 and Figure F-4 present the proportion of cracks exceeding crack sizes for increasing median size at a constant $\sigma = 0.75$ and increasing sigma at a constant crack size median of 10 mils. The $POD(a)$ function with $a_{50} = 50$ and $\sigma = 0.5$ is again included as the size reference. Note that under the assumed scenarios, relatively few of the cracks will have sizes that would be in a range with POD greater than 0.9.

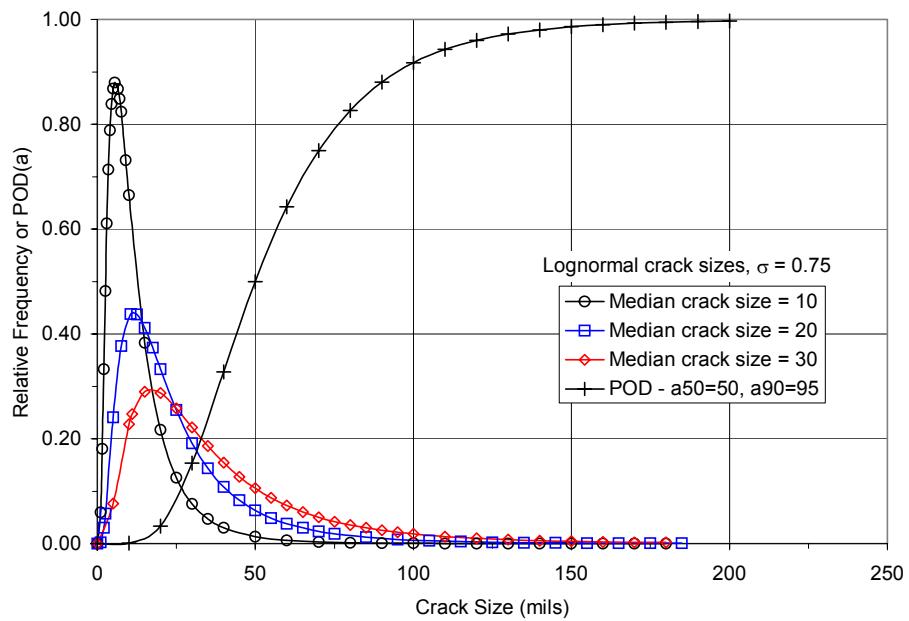


Figure F-2: Probability Density Functions of Crack Size Distributions with POD(a) for $\sigma = 0.5$.

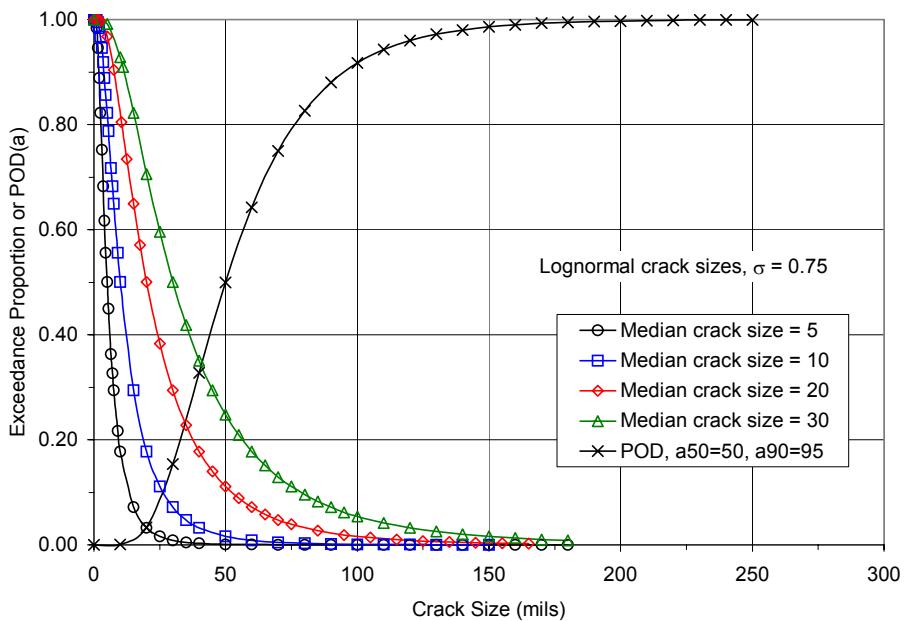


Figure F-3: Exceedance Probabilities of Crack Sizes for Increasing Median Size and $\sigma = 0.5$.

ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS

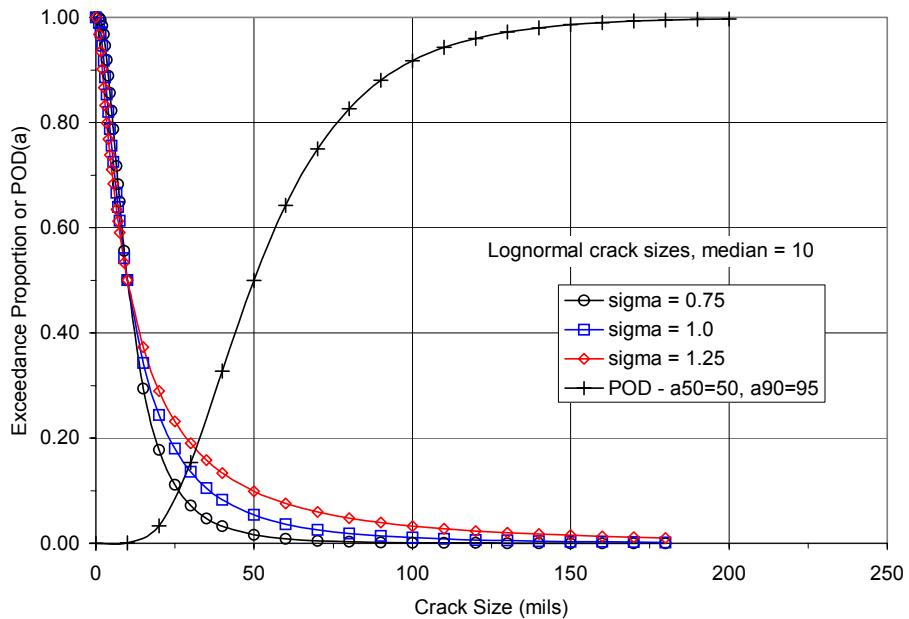


Figure F-4: Exceedance Probabilities of Crack Sizes for Increasing σ with Median = 10 mils.

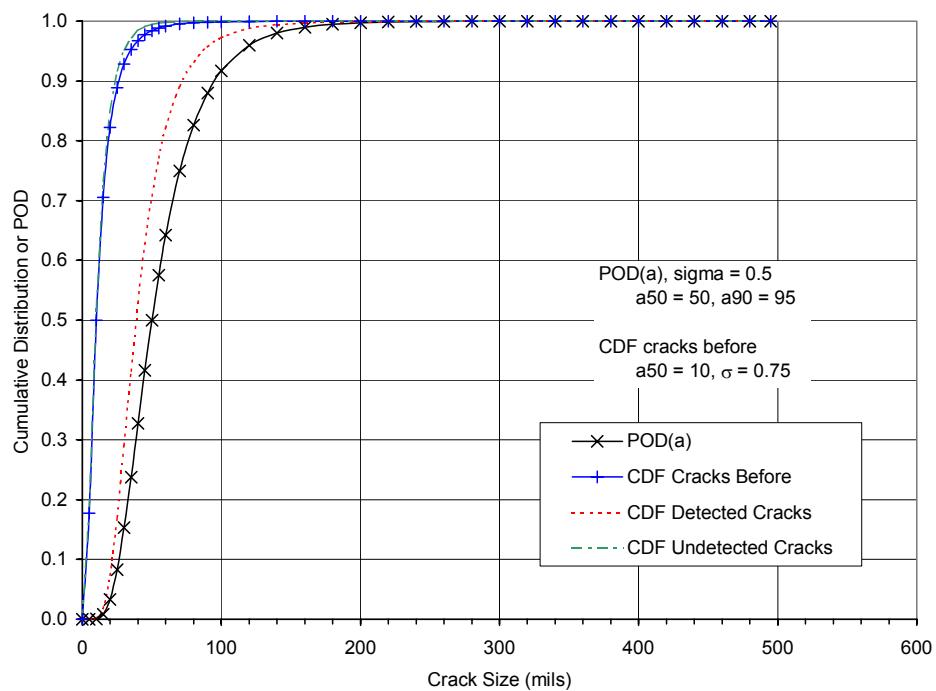
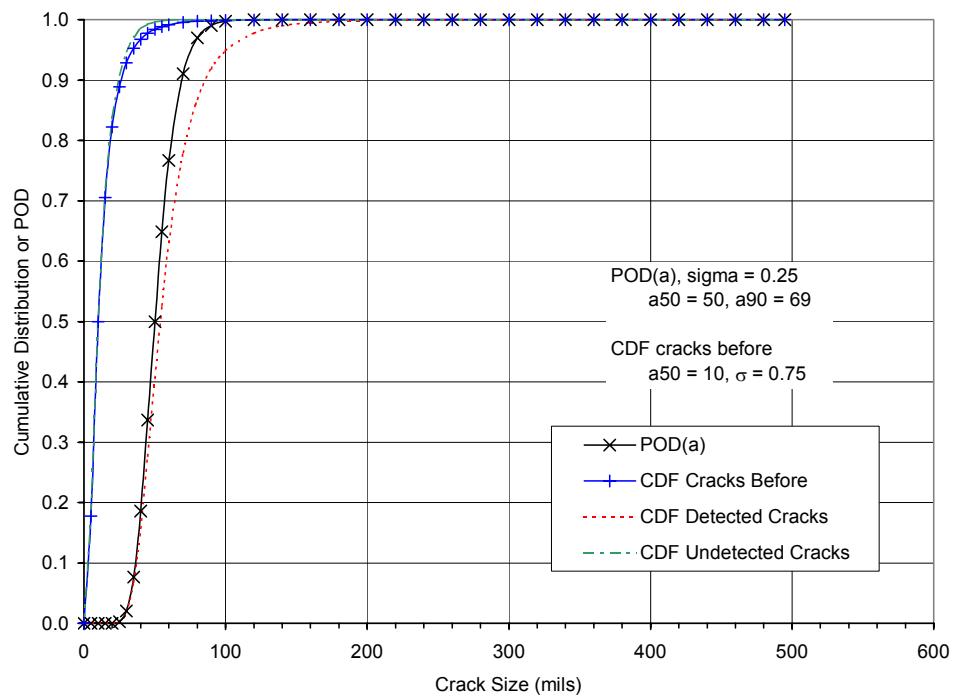
Table F-1 presents the total proportion of the cracks that would be detected for each of the 60 combinations of $POD(a)$ and crack size being considered. These proportions can be interpreted in the context of the inspections of the F-16 center fuselage longeron. In these inspections, 39 cracks were detected in a minimum of 280 inspections. A maximum of 14% of the inspections resulted in detection. The combinations of crack size and $POD(a)$ that are inconsistent with the in-service inspections (i.e. exceed 14%) are shaded in Table F-1.

Table F-1: The Proportion of Inspections that Result in Detection

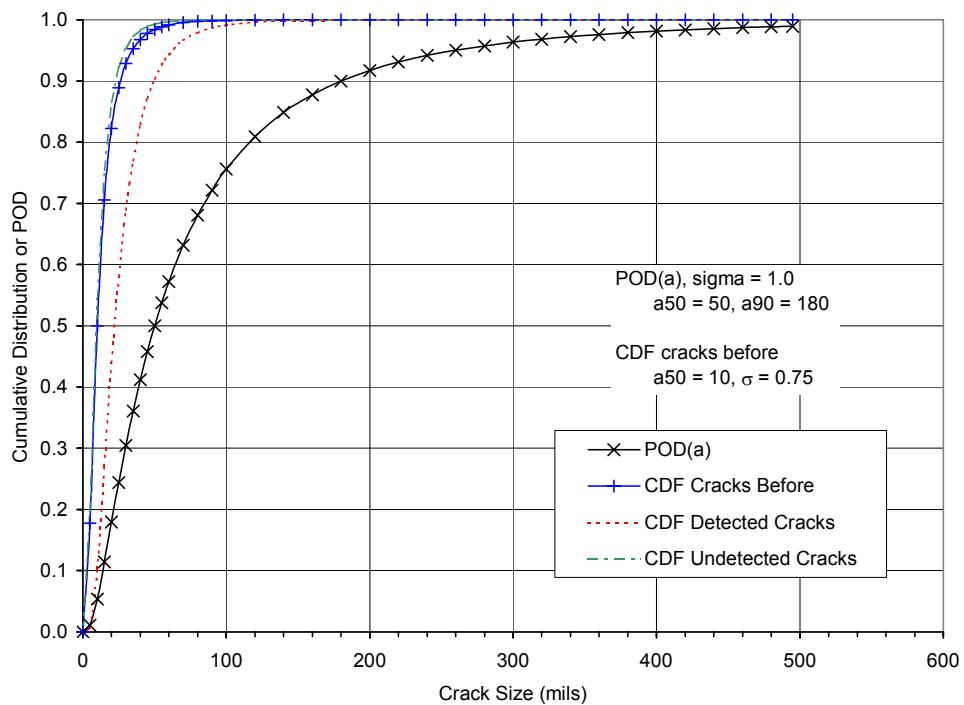
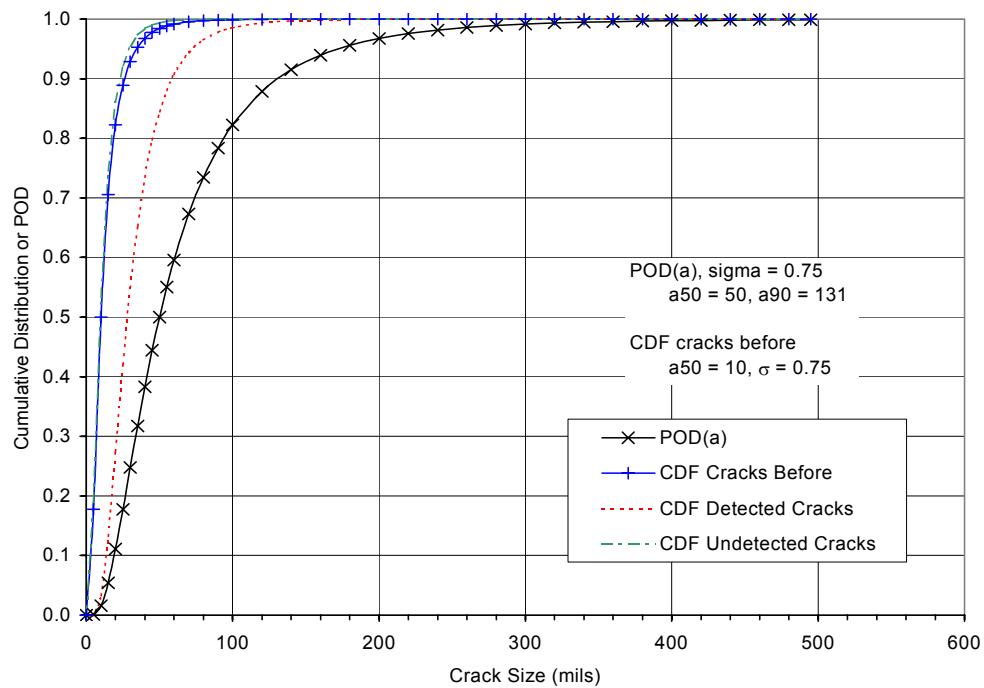
Log-Normal Crack Sizes		Log-Normal $POD(a) - POD(50) = 0.50$ σ for $POD(a)$				
σ	a_{50}	0.25	0.50	0.75	1.00	1.25
0.75	5	0.002	0.005	0.015	0.033	0.057
	10	0.021	0.037	0.065	0.099	0.135
	20	0.123	0.155	0.194	0.232	0.265
	30	0.259	0.285	0.315	0.341	0.363
1.00	5	0.013	0.020	0.033	0.052	0.075
	10	0.059	0.075	0.099	0.128	0.157
	20	0.187	0.206	0.232	0.259	0.284
	30	0.310	0.324	0.341	0.359	0.375
1.25	5	0.035	0.044	0.057	0.075	0.096
	10	0.103	0.116	0.135	0.157	0.181
	20	0.236	0.248	0.265	0.284	0.302
	30	0.344	0.352	0.363	0.375	0.386

Restrict attention to the crack size distribution with a median size of 10 mils and a standard deviation of 0.75. For the five NDI capabilities, Figure F-5 through to Figure F-9 show the $POD(a)$ function, the distribution of crack sizes before the inspection, $F(a)$, the distribution of the sizes of cracks detected during the inspection, $G_{det}(a)$, and the distribution of the sizes of cracks that were not detected during the inspection, $H_{miss}(a)$. As σ of $POD(a)$ increases, the distribution of detected cracks shifts to smaller sizes. The increasing proportion of total inspections that result in a detection, as listed in Table F-1, is due to the increasing number of smaller cracks that are detected by the greater $POD(a)$ capability at the smaller sizes. Stated in terms of the reliably detected crack size, the larger the a_{90} , the smaller are the cracks that will be detected.

ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS



ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS



ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS

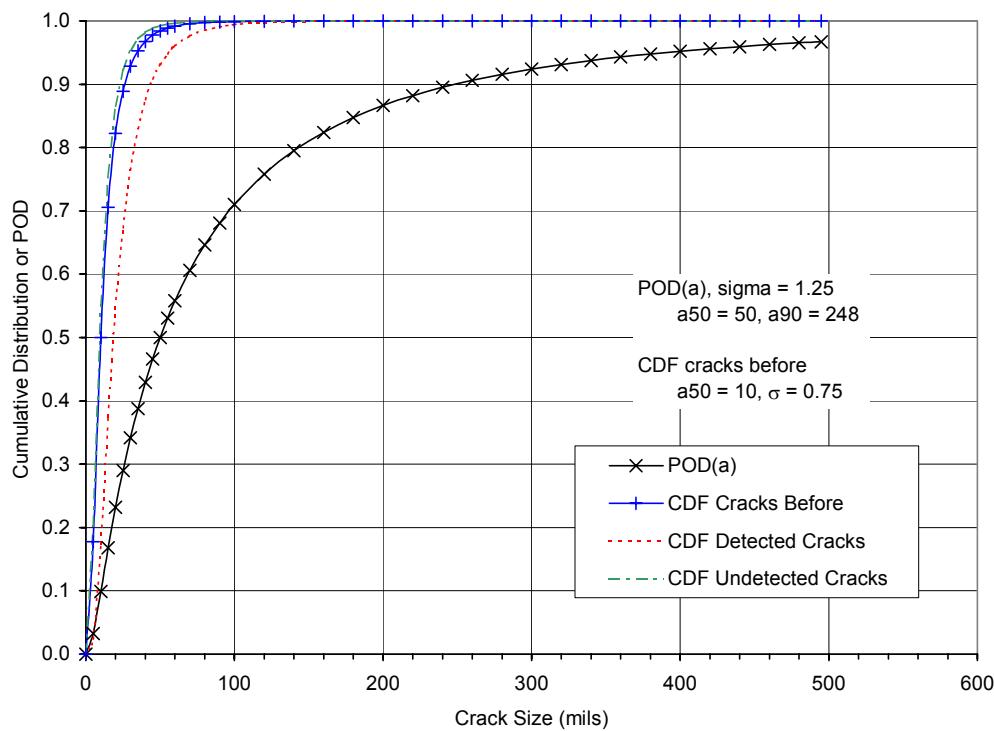


Figure F-9: Detected and Undetected Crack Sizes:
POD(a) – $a_{50} = 50$, $\sigma = 1.25$, Initial Crack Sizes: Log-Normal – $a_{50} = 10$, $\sigma = 0.75$.

Figure F-10 and Figure F-11 present the same general results for crack sizes with a median of 5 mils and a standard deviation of 1.25. Figure F-10 represents the best inspection capability and the most scatter in crack sizes. In Figure F-10, $\sigma = 0.25$, and the a_{90} value for the POD(a) function is 69 mils. This capability would be considered as excellent. About 3.5% of the inspections would result in crack detection, but only relatively large cracks would be detected. The 90th percentile of the detected cracks is about 145 mils. Figure F-11 represents the worst inspection capability ($a_{90} = 248$ mils) and the most scatter in cracks sizes. For this combination, about 10% of the inspections will result in detection. Many smaller cracks would be detected because POD(a) is significantly greater over the range of crack sizes. The 90th percentile of the detected cracks from this scenario is about 80 mils.

ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS

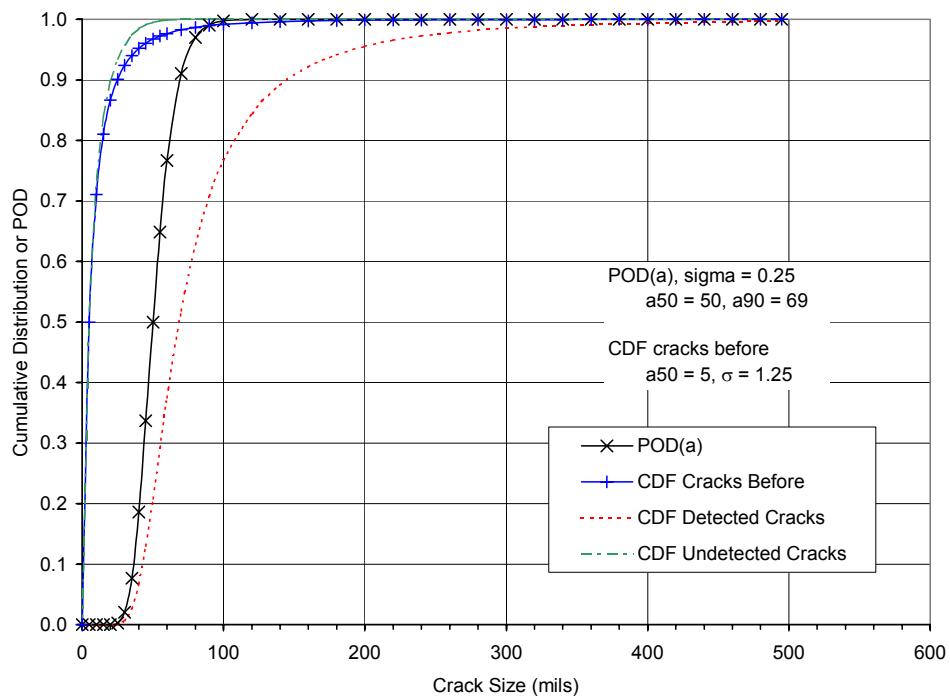


Figure F-10: Detected and Undetected Crack Sizes:
POD(a) – $a_{50} = 50$, $\sigma = 0.25$, Initial Crack Sizes: Log-Normal – $a_{50} = 5$, $\sigma = 1.25$.

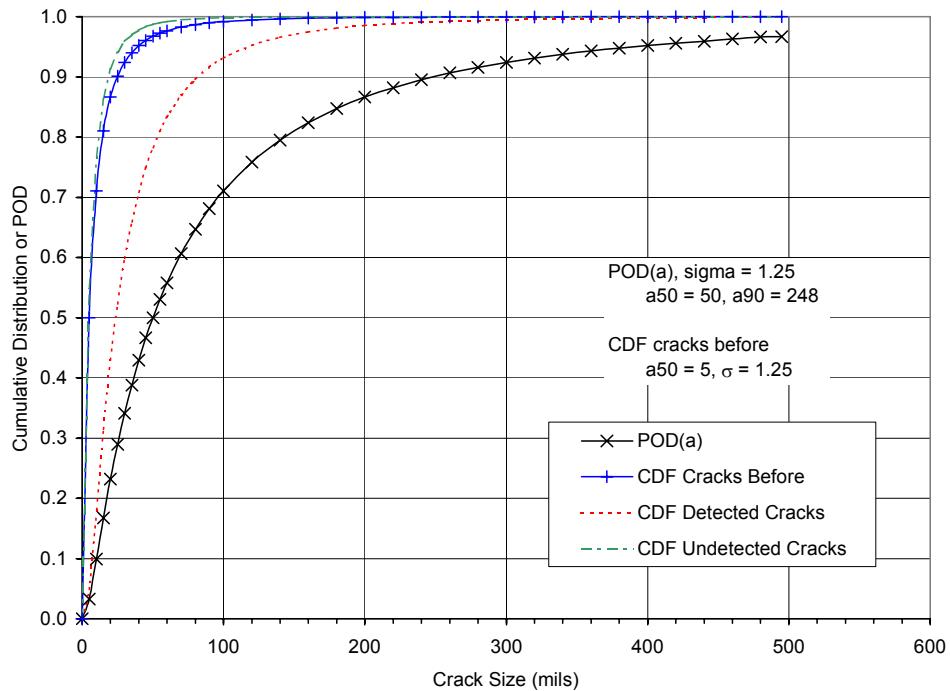


Figure F-11: Detected and Undetected Crack Sizes:
POD(a) – $a_{50} = 50$, $\sigma = 1.25$, Initial Crack Sizes: Log-Normal – $a_{50} = 5$, $\sigma = 1.25$.

ANNEX F – DISTRIBUTIONS OF DETECTED AND UNDETECTED CRACKS



Annex G – CRACK SIZE ERRORS – NATURE AND EFFECT ON POD ESTIMATION

G.1 MATERIAL DIFFERENCE EFFECT ON THE VARIABILITY OF BACK-CALCULATED CRACK SIZES

To obtain a lower bound on the amount of scatter that might result from the back-calculation of crack sizes at previous times, actual crack growth data from 68 replicate tests were analyzed. The test program, conducted by Virkler and Hillberry at Purdue University, is documented in Virkler et al. (1978). Sixty eight (68) identical 2024-T3 aluminum center cracked panels were cycled under constant amplitude loading until failure. The panels were 25.4 mm (0.1 inch) thick and 152 mm (6.0 inch) wide. The maximum load was 23.4 kN (5.25 KIP) with a stress ratio of 0.2. Crack growth as a function of cycles was determined by recording the number of cycles required at each 0.2 mm of crack growth. All time histories were translated to an initial size of 9 mm at zero cycles. Figure G-1 presents a plot of crack size versus cycles for all 68 test specimens. The amount of scatter exhibited in Figure G-1 is due to material properties. Crack growth models cannot account for individual deviations from the average.

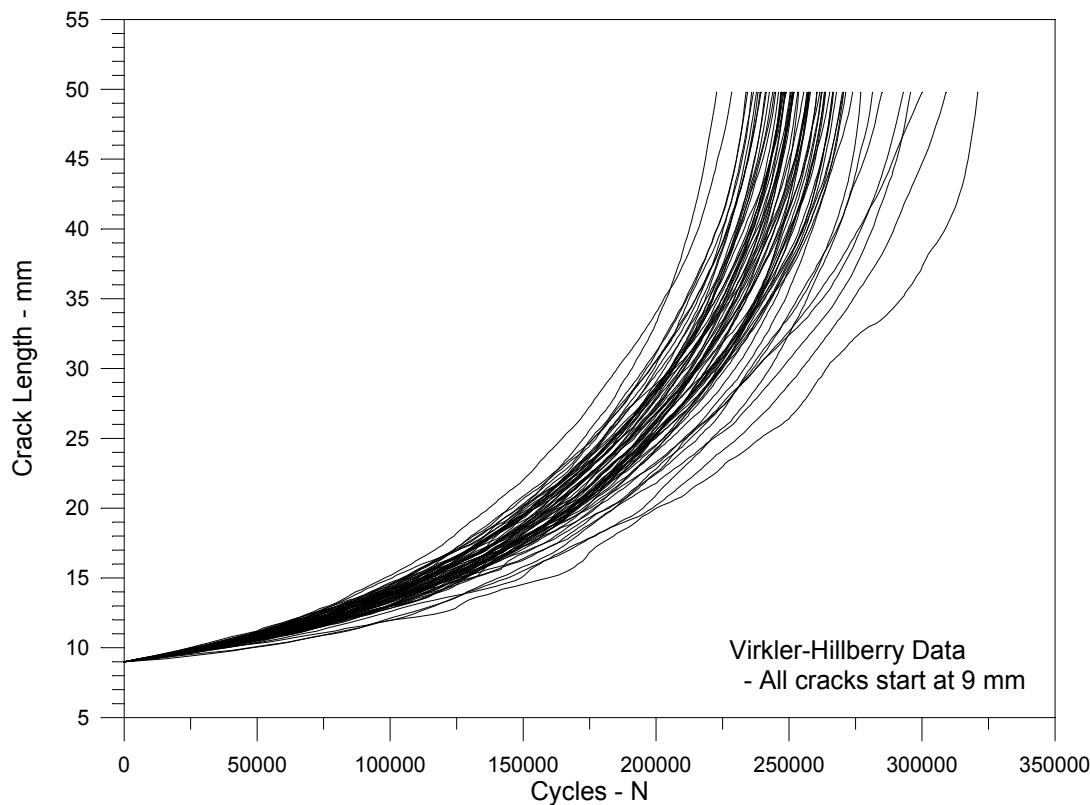


Figure G-1: Crack Growth versus Constant Amplitude Cycles for 68 Identical Tests.

In the POD scenario of this study, a crack size is observed at a given life, and the size at a previous point in time is calculated. To investigate the material difference effect on scatter in crack sizes at a previous time, a crack size and corresponding life were selected from the mean crack size curve of the 68 specimens of

Figure G-1. Figure G-2 presents the average crack size as a function of cycles. Each of the 68 crack growth curves were then translated to pass through this fixed size and number of cycles. The shape of the individual histories was not changed. The distribution of crack sizes at previous fixed points in time were then read from the translated curves.

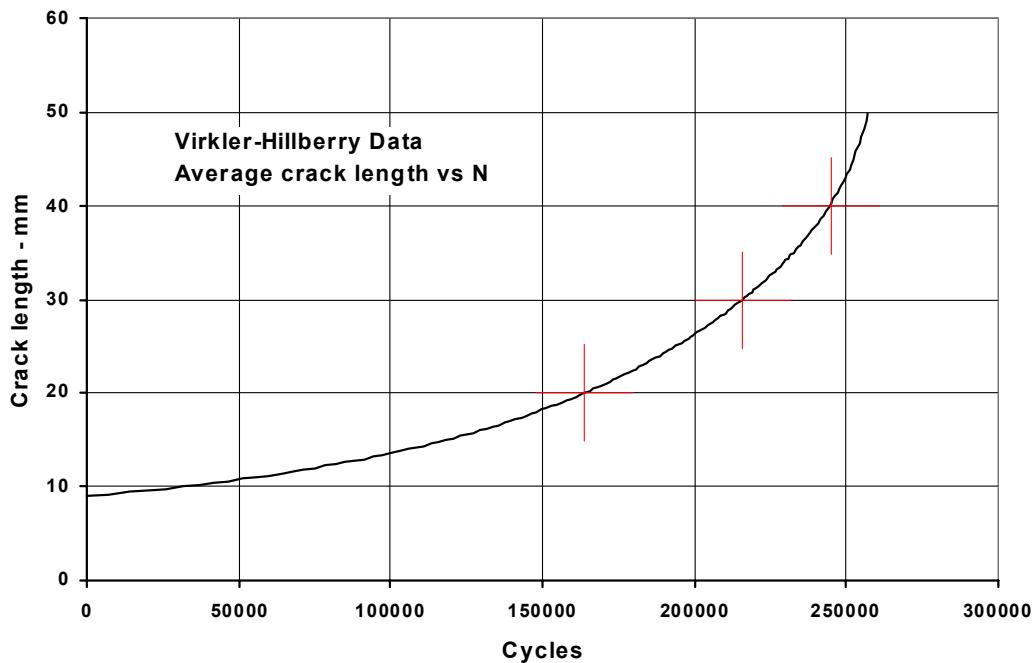


Figure G-2: Average Crack Size versus Cycles.

Initial crack size starting points were selected at 20, 30 and 40 mm at lives of 164,000, 216,000 and 245,000 cycles, respectively. These points are indicated on Figure G-2. The original crack growth histories were translated horizontally to pass through each of these points with the resulting crack curves as shown in Figure G-3 through Figure G-5. The crack sizes at the indicated cyclic lives of 66,000, 116,000 and 166,000 were then interpolated from each of the 68 specimen histories. The scatter in these sizes is indicative of the crack size errors that could result from only material differences in a back-calculation over periods of about 50,000, 100,000 and 150,000 cycles (about 20, 40 and 60 % of the average specimen life) starting at three different crack sizes.

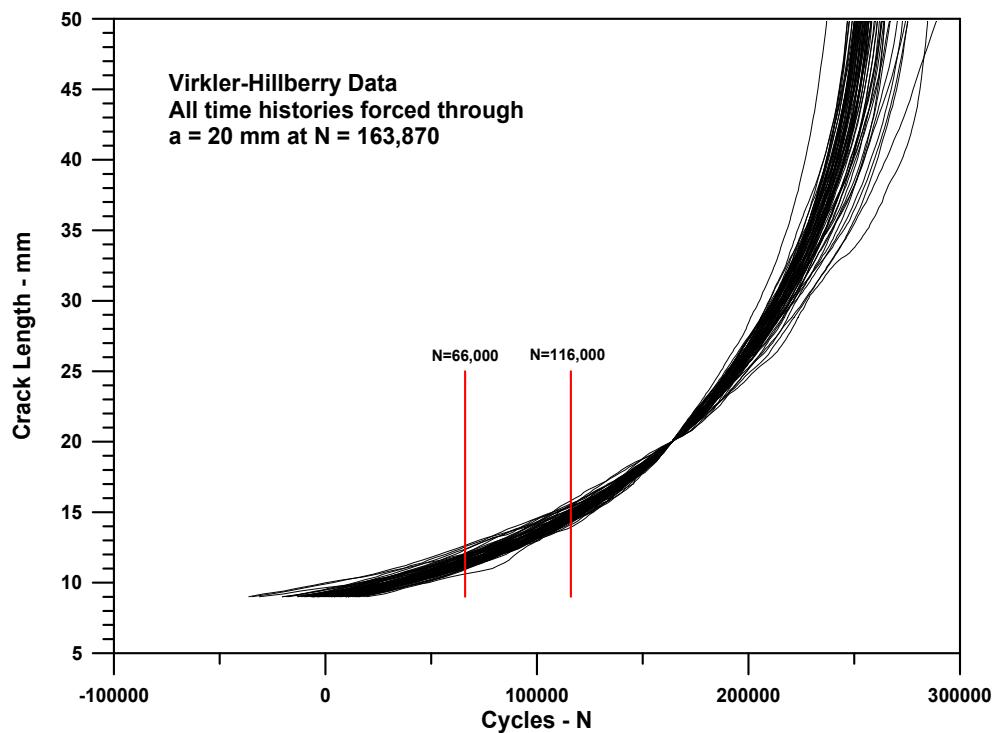


Figure G-3: Crack Growth Histories Coincident at 20 mm.

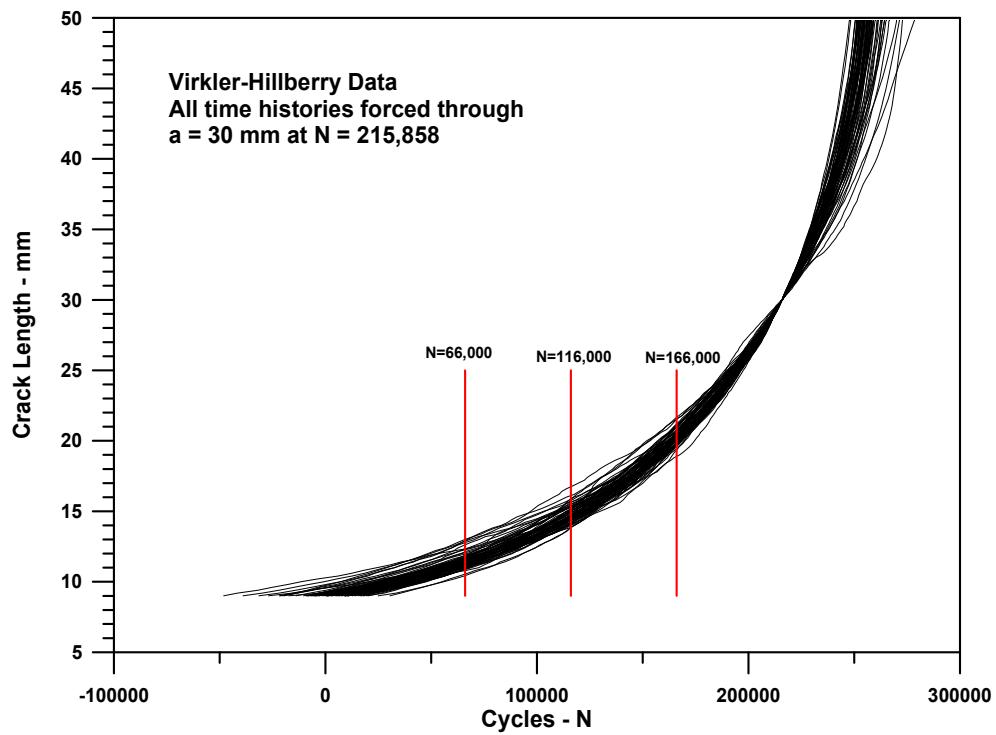


Figure G-4: Crack Growth Histories Coincident at 30 mm.

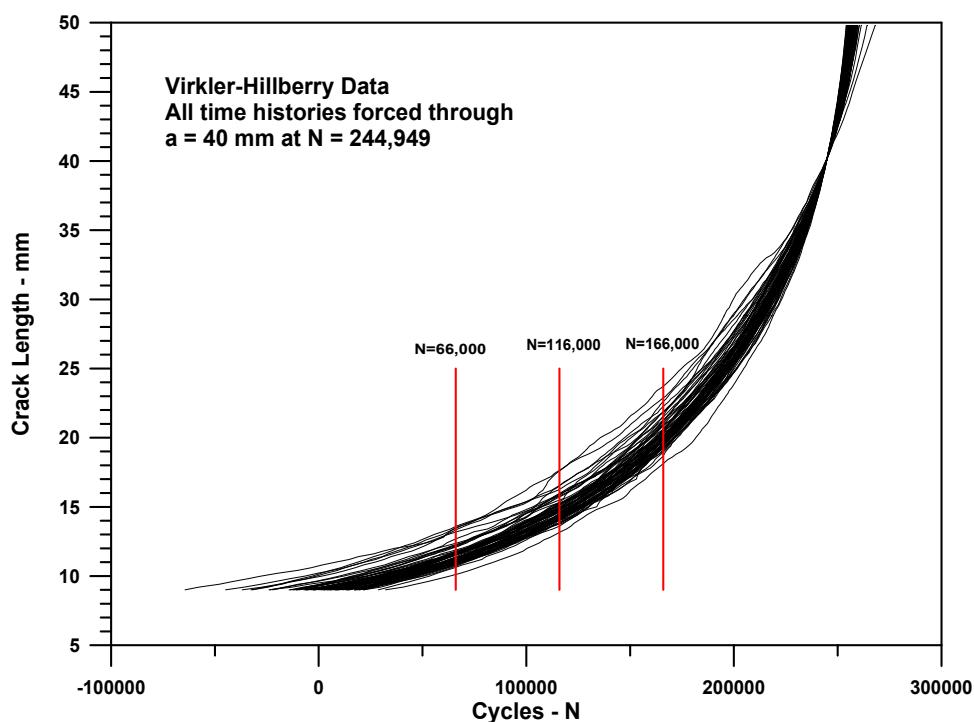


Figure G-5: Crack Growth Histories Coincident at 40 mm.

The averages, standard deviations and coefficients of variation of the back-calculated crack sizes at the indicated cyclic lives of 66000, 116000 and 166000 cycles are shown in Table G-1. For reference, the same statistics from the crack sizes from the original data are also included. Figure G-6 presents the cumulative distributions of the back-calculated crack sizes.

Table G-1: Summary Statistics of Crack Sizes from Back-Calculations from 68 Actual Crack Growth Histories

Averages of Back Calculated Crack Sizes - mm			
	To N=66,000	To N=116,000	To N=166,000
From a=20, N=163,870	11.52	14.79	
From a=30, N=215,858	11.52	14.78	20.27
From a=40, N=244,949	11.52	14.78	20.28
From a=9, N=0	11.56	14.88	20.45
Standard deviations of Back Calculated Crack Sizes - mm			
	To N=66,000	To N=116,000	To N=166,000
From a=20, N=163,870	0.409	0.387	
From a=30, N=215,858	0.535	0.585	0.541
From a=40, N=244,949	0.702	0.813	1.010
From a=9, N=0	0.381	0.838	1.743
Coefficients of Variation of Back Calculated Crack Sizes			
	To N=66,000	To N=116,000	To N=166,000
From a=20, N=163,870	3.6%	2.6%	
From a=30, N=215,858	4.6%	4.0%	2.7%
From a=40, N=244,949	6.1%	5.5%	5.0%
From a=9, N=0	3.3%	5.6%	8.5%

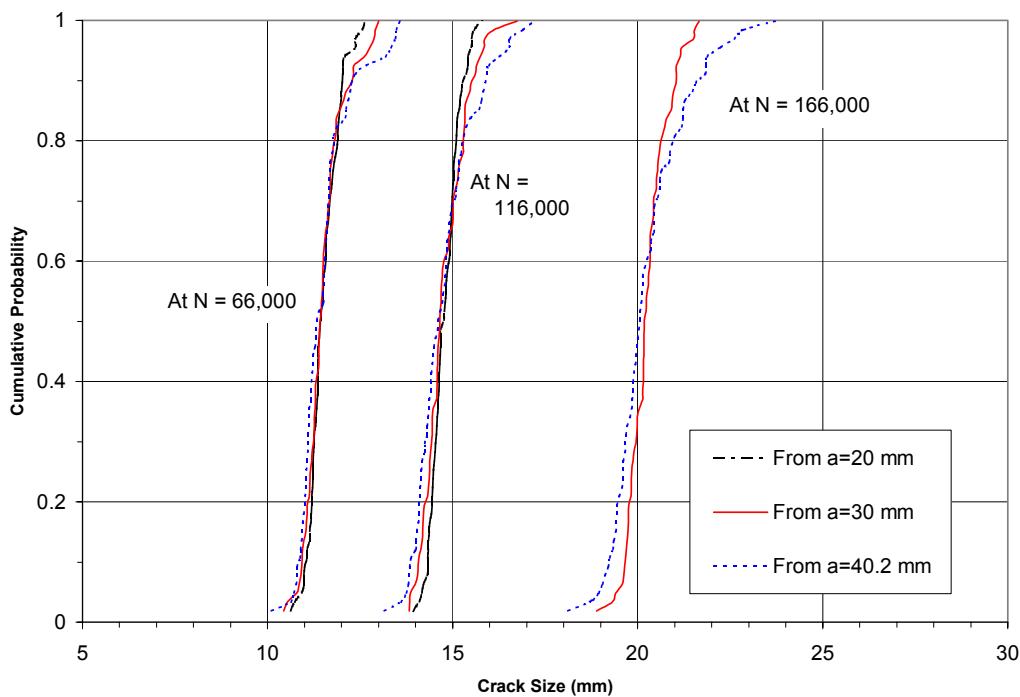


Figure G-6: Distributions of Back-Calculated Crack Sizes at Selected Times.

The mean crack sizes of the back-calculated lives are essentially equal at each of the three cyclic times. This may be due to the method of back-calculation using the time histories. The scatter about the means is consistent with the scatter in the original time histories of crack growth in which the coefficient of variation of crack size increases with experienced load cycles. The coefficient of variation of the back-calculated crack sizes increases with the length of the period of back-calculation, but is equivalent to that of the original data.

The crack growth histories from these identical tests of aluminum panels suggest that the minimum scatter in back-calculated crack sizes would be of the order of a 5% coefficient of variation. Because the loads experienced during operational experience are not precisely known and the back-crack size estimation must be performed analytically, the degree of scatter in the real application could be significantly greater. However, the effect of at least this degree of variability on an estimate of POD should be determined.

G.2 CRACK LENGTH ERROR EFFECT ON POD ESTIMATION (VIA REGRESSION CURVES)

The standard models (“hit/miss” and \hat{a}) used for POD estimation are regression models. In the regression models, the crack length is the independent variable. “The study of regression models wherein the independent variables are measured with error predates the twentieth century.¹” Here, a brief development is given in terms of POD models. Fuller (Fuller, W.A. (1987), *Measurement Error Models*, John Wiley & Sons, New York) considers the subject in great detail.

¹ Opening sentence in the Preface of Fuller (1987).

In the following, the symbol “ s ” (for signal) is used instead of \hat{a} and the variable x is the crack length. The basic regression model is given by modeling the relation of the dependent variable, s , to the independent variable x :

$$s = c + d \cdot \ln(x) + \varepsilon, \quad \varepsilon \approx N(0, \sigma_\varepsilon^2) \quad (\text{G-1})$$

That is, there is a mean relationship between the signal and the flaw length plus a random noise variation in the signal. It is usual to assume that the noise has a Gaussian distribution with 0 mean and variance, σ_ε^2 . (The source of this noise is crack-to-crack variations, as well as implementation and instrument noise.) Of course, it is understood that the signal, s , and the crack length, x , in the model could be transformations. The model is given here using the logarithm transform on the crack size, as this is the usual model.

If the relationship in (G-1) holds and a threshold, T , is established for the signal to give an indication during an inspection, then the POD is determined by $POD(x) = \Pr(s > T | x) = \Phi\left(\left[\ln(x) - \frac{T-c}{d}\right] / \frac{\sigma_\varepsilon}{d}\right)$, where Φ is the standard normal distribution function. The last functional form emphasizes the POD as equivalent to the distribution function for a log-normal random variable with parameters, $\mu = \frac{T-c}{d}$, and standard deviation, $\gamma = \frac{\sigma_\varepsilon}{d}$. It is the parameters μ and γ that are estimated directly in a “hit/miss” analysis.

The development presented here depends on the μ and γ parameters by the above transforms and apply equally to “hit/miss” analysis and to an \hat{a} analysis.

The two sources of measurement error modeled are a fixed bias, as well as random noise in the measurement. That is,

$$\ln(x') = b + \ln(x) + \delta, \quad \delta \approx N(0, \sigma_\delta^2), \quad (\text{G-2})$$

where b is the bias in the log-scale and x is the true, but unknown crack length, and x' is the crack length used in regression. In this formulation, the bias, b , and the random error δ are both relative errors in the original measurement scale.

Substituting (G-2) into (G-1) the model that would be considered in regression can be expressed

$$s = c^* + d \cdot \ln(x') + \varepsilon^*, \quad \text{where } c^* = c - d \cdot b \text{ and } \varepsilon^* = \varepsilon - d \cdot \delta. \quad (\text{G-3})$$

The usual regression analyses gives estimates for c^* , d and variance, ε^* . The POD analysis estimates parameters μ and γ , as given above. First the effect of the measurement error on the regression parameters will be discussed and then the transformations to the POD parameters will be discussed.

The effect of the slope, d , of the signal to log-flaw size is confounded with the intercept as well as the residual term when measurement error is present. The derivation is not given here, but it can be shown that for the model of equation (G-3) the usual regression estimate of the slope parameter, d , is biased and the expectation is given by equation (G-4)

$$E(\hat{d}) = d \cdot \frac{(1 + \rho \cdot r)}{(1 + 2\rho \cdot r + r^2)}, \quad (G-4)$$

where, letting $\sigma_{\varepsilon\delta} = \text{covariance}(\varepsilon, \delta)$, then the additional parameters of equation (G-4) are given by $\rho = \sigma_{\varepsilon\delta}/(\sigma_\varepsilon \cdot \sigma_\delta)$ and $r = \sigma_\delta/\sigma_\varepsilon$. If the measurement system for a crack was independent of the non-destructive technique used to detect the crack, then $\sigma_{\varepsilon\delta} = 0$ and thus $\rho = 0$ and $E(\hat{d}) = d/(1 + r^2) = d \cdot \sigma_\varepsilon^2 / (\sigma_\delta^2 + \sigma_\varepsilon^2)$. This is a well-known result in the regression literature where it is said that the regression coefficient has been *attenuated* by the measurement error.

The bias factor in equation (G-4) is less than 1 as long as $r + \rho > 0$ or $\sigma_\delta^2 + \sigma_{\varepsilon\delta} > 0$, which is likely to always be the case in applications. The crack length measurement may be uncorrelated with the signal ($\rho = 0$), but since flaw length may be determined after a signal is obtained, there may be a tendency to estimate flaw length in the same direction as the signal (i.e. $\rho > 0$). This would be especially true if NDE was used in crack sizing.

Returning to the POD estimate, what is the effect on the estimates of the parameters $\mu = (T - c)/d$ and $\gamma^2 = \sigma_\varepsilon^2/d^2$? (It is more natural to consider the variance parameter, γ^2 , rather than the standard deviation γ .) In the regression, $d^2 \cdot \gamma^2 = \sigma_\varepsilon^2$ and σ_ε^2 would be estimated from the residuals in equation (G-3), $\varepsilon^* = \varepsilon - d \cdot \delta$. These residuals have mean 0 and variance $\sigma_\varepsilon^2 - 2d \cdot \sigma_{\varepsilon\delta} + d^2 \cdot \sigma_\delta^2$. Therefore the variance of $\hat{\gamma}$ is given by

$$E[\hat{\gamma}^2] = \frac{\sigma_\varepsilon^2 - 2 \cdot d \cdot \sigma_{\varepsilon\delta} + d^2 \cdot \sigma_\delta^2}{d^2} = \gamma^2 + \sigma_\delta^2 - \frac{2\sigma_{\varepsilon\delta}}{d} \quad (G-5)$$

NOTE: Equation (G-5) assumes that an unbiased estimator for σ_ε^2 is used in a residual analysis. Using the maximum likelihood estimator, which is not unbiased, requires the added factor of $(n - 1)/n$, where n is the number of cracks.

The mean of the POD function with measurement error is given by

$$E[\hat{\mu}] = E[(T - c^*)/d] = E[(T - c)/d + b] = \mu + b \quad (G-6)$$

Therefore, the crack length bias is reflected by a direct shift of the mean of the POD and the variance parameter is increased by the variance of the measurement error with an adjustment that is dependent upon whether the measurement errors are correlated with signal size. Because the formulation of the problem was in terms of the logarithm of crack length, σ_δ is interpreted as the relative error. Therefore, in the previous section, the 5% coefficient of variation translates to $\sigma_\delta \cong 0.05$.

The effect of crack length measurement error is relative to the POD estimated parameters. The general advice in regression problems is that errors in the independent variable can be ignored if the variance in measurement is “small” compared to the residual variance of the response variable. The same criterion applies here, where the POD-scale parameter serves in the role of the response variable error, as can be seen in equation (G-5).

ANNEX G – CRACK SIZE ERRORS – NATURE AND EFFECT ON POD ESTIMATION

Based on the regression models that are used to estimate POD curves, the above derivations give an indication of the impact of measurement errors made on the independent variable, crack size when that error applies to all the crack length measurements. However, the crack lengths used in calculating a POD curve are those that have been measured in the field (the hits) plus those crack lengths that have been inferred from back-calculations using an average crack growth curve. In this case, the “hits” and “misses” have different error sources for the crack lengths. The “hits” are subject to the field measurement errors, but the crack lengths for the “misses” are subject to: 1) the measurement error in the crack from which the miss lengths are inferred; 2) the possible error in the choice of appropriate crack growth mean line; and 3) the natural flaw variation discussed in previous section. It is therefore likely that the overall variation in the flaw lengths used in POD estimation associated with the “misses” will be greater than that associated with the “hits”. These sources of error will also contribute both a random component as well as bias components. It is likely that the bias component will impact the uncertainty through equation (G-6) with more impact than the random component.

Annex H – BINOMIAL AND BAYESIAN ANALYSIS

H.1 RATIONALE

H.1.1 Limitations of Probabilities of Detection and Safety Levels Deduced from Small Samples

Whenever NDT is used on aircraft primary structure to detect potentially critical defects, the reliability of the inspection method used becomes one of the principal factors determining the safety level at which the aircraft operates. A common rule of thumb, although not used by the USAF, assumes that a defect should be inspected at least three times during the period in which it grows to the maximum acceptable size. If the probability of detecting the defect is 90% for each of the three inspections, this leads to a safety level, the probability of missing the defect completely, of 1 in 1000. This level is similar to the 1 in 1000 probability of failure due to fatigue crack growth on which safety factors for safe-life airworthiness assessment are frequently based.

The standard USAF methodology (see the main report, Section 5.2) for assessing inspection reliability, characterising the inspection process by a 95% confidence level POD curve estimated from artificial trials, has become the standard approach. In the particular case of POD analyses carried out for the USAF in support of damage tolerance-based life assessment, the parameter used to characterise an inspection is the “detectable” crack size $a_{90/95}$. This is defined as the minimum crack length at which a 90% POD has been demonstrated at the 95% confidence level. This method works satisfactorily for straightforward inspection situations where the POD curve can be estimated from a large database, such as occurs in engine disk inspection for example. Application of similar methods to airframe inspection suffers from the prohibitive cost of obtaining the reliability curve from realistic trials. Where there is only a limited amount of data available to determine the reliability of an inspection method, the in-built conservatism of the standard method may lead to unrealistic estimates for the 95% confidence POD curve or $a_{90/95}$ value. This may in turn give rise to unacceptably short inspection intervals and excessive maintenance costs.

In order to overcome the problem of providing realistic data, this Working Group study has looked at whether it is practicable to estimate inspection reliability from in-service inspection data. Although many inspections are carried out and many defects found, the diversity of inspection situations including access, geometry and equipment variations suggest that there will still be a very limited amount of information available from which to estimate the reliability for many inspection tasks. If the NDE inspection results are likely to be insufficient to validate the standard 90% POD at 95% confidence requirement for a specified critical crack size, it is necessary to assess whether there is a better way of measuring and reporting NDE reliability. Viewed from the NDE perspective, the need is for a statistical method which will most efficiently make use of whatever data can be collected to predict the probable outcome of future inspections.

H.1.2 Single Probabilities of Detection for Homogeneous Defects

The simplest case can be thought of as the task of predicting the probability of missing a defect during the number of inspections, typically three or so, which will be carried out in service during the defect growth phase, assuming that the probability of detection is constant. The most straightforward approaches take the form of trying to predict as accurately as possible the probability of an expected outcome. The standard method of analysing NDT reliability, establishing a lower bound and then using this lower bound to estimate the probability of missing a defect three times, say, is unusually conservative.

The extent of the conservatism in the estimates incorporated in the standard methodology can be illustrated by estimating the probability of missing a defect three times, given that the required POD(a) of 0.9 at 95% has been verified. The conservative estimate is simply obtained by assuming that the true probability is equal to the lower bound of 0.9, in which case the probability of three misses is simply 0.001. In reality, in order to verify the POD(a) value, the actual value for the technique must be significantly higher than 0.90. Using the mean probability, p_m , to estimate the outcome of the three inspections leads to the values in Table H-1, where the results of the initial verification exercise are given, together with the resulting p_m and the most likely prediction for probability of three misses.

Table H-1: Estimated Safety Level, i.e. Probability of Three Successive Misses, after Verifying a POD of 0.90 at 95% Confidence using Minimum Sample Sizes

Initial experiment		Probabilities		POD = 0.90
			Prob of 3 misses	
Hits	Trials	p_m	$(1-p_m)^3$	$(1-p_a)^3$
29	29	1	0	0.001
45	46	0.978	1.03E-05	0.001
59	61	0.967	3.52E-05	0.001
73	76	0.961	6.15E-05	0.001
85	89	0.955	9.08E-05	0.001
98	103	0.951	0.000114	0.001
122	129	0.946	0.00016	0.001
157	167	0.940	0.000215	0.001

It can be seen that the likely performance of the technique is very much better, possibly one or two orders of magnitude better than the conservative estimate predicts. This degree of conservatism is acceptable if sufficient information is available to verify the high POD value, however, it is a luxury if it is unrealistic to expect the limited data available to provide such high estimates. In-service data is quite likely to yield lower estimates. For example, the best information available to the WG came from The Netherlands and Canadian Air Forces who reported maximum defect numbers of 39 and 25, respectively, for specific inspections.

There are various methods of predicting the probability or likelihood of an outcome based on an initial experiment. The most straightforward are based on the use of a contingency table and a standard statistical test such as the χ^2 or Fisher's likelihood test. These approaches allow the probability of missing a defect three times after the initial experimental result to be deduced directly, without recourse to calculating an intermediate POD for a single trial.

A more elegant method can be based on Bayesian inference. In the Bayesian approach, the degree of confidence in a particular outcome before an experiment is expressed as a "prior" distribution of probabilities. In applying the approach to NDT reliability assessment, the prior distribution is chosen as the level of confidence in achieving given values for the probability of detection. An initial experiment is then carried out. The outcome of the initial experiment is used to update the prior distribution, producing a "posterior" distribution reflecting the revised degree of confidence in the possible outcomes as a result of including the additional information which has been obtained. Bayesian confidence levels and intervals can be estimated from the posterior distribution which can be used to determine the effectiveness of the inspection. Finally, the Bayesian analysis can be used to produce a third distribution, the "predictive" distribution, which is calculated

directly from the posterior distribution. This gives the probability of any outcome in a subsequent experiment given the initial level of knowledge in the prior distribution and the additional information from the initial experiment.

A useful concept in Bayesian analysis is the use of conjugate pairs of distributions. The results of the experiments can be described by one type of distribution, in this case the binomial distribution. If a prior distribution can be chosen from a family of distributions so that the posterior distribution calculated from the experiment is from the same family as the prior distribution, then the two distribution types are said to be conjugate. Since the prior and posterior are of the same type, it follows that any further experiments can be used to generate a further posterior distribution incorporating all of the experimental information which will again belong to the same family of distributions. In the case of the binomial distribution $p(h,n,p_0)$, it is known that the conjugate distribution is the Beta distribution $Be(\gamma, \eta, p)$, where γ and η are constants. The predictive distribution formed from the Beta distribution is called the Beta-Binomial distribution, $BeBi(h_2, n_2, \gamma, \eta)$, where h_2 and n_2 are the assumed hits and trials in the subsequent experiment.

The prescription for analysing reliability experiments in this formalism is then to start with a prior distribution from the Beta family. An initial experiment or a series of inspections in service will provide a known number of hits and misses which can be used to update the prior. It can be shown that if the prior is $Be(\gamma, \eta, p)$ and a binomial experiment has resulted in h hits and $n - h$ misses, then the resulting posterior distribution is $Be(\gamma+h, \eta+n-h, p)$ and the predictive distribution is $BeBi(h_2, n_2, \gamma+h, \eta+n-h)$. The results of subsequent experiments or periods of inspections in service can naturally be built into the posterior distribution by using the total numbers of “hits” and “misses” to date.

The safety level can be calculated from these probabilities p_m and $POD(a) (\equiv p_\alpha)$ and from the Bayesian predictive Beta-Binomial distribution. If it is assumed that three inspections will be carried out in service on the defects, the appropriate expressions are:

$$\text{Binomial; } p(0, 3, p_{\alpha/m}) = (1 - p_{\alpha/m})^3$$

$$\text{Bayesian; } p(0, 3, n) = BeBi(0, 3, 1+h, 1+n-h)$$

H.1.3 Examples

The process can be illustrated by simulating a reliability verification experiment carried out in small sets of trials. This would provide data similar to the slow accumulation of real data which might be expected from the results of in-service inspection. In the example below, it is assumed that the underlying probability of detection, the true probability, is 0.92. A total of 45 inspections have been carried out in groups of 5 inspections. An initial prior distribution has been chosen with $\gamma = \eta = 1$, which gives a uniform distribution indicating that no information on the reliability of the technique is available. Figure H-1 shows the posterior distributions after each set of 5 trials.

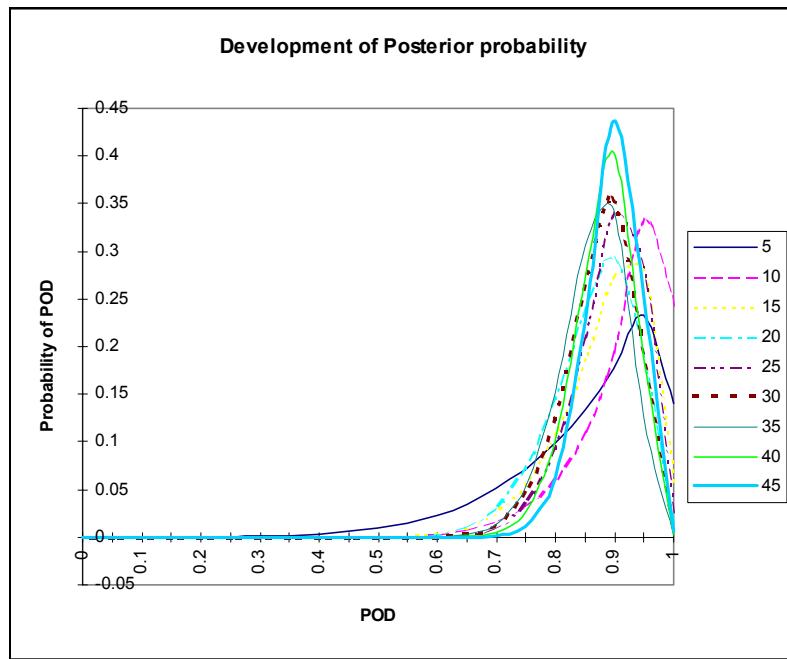


Figure H-1: Evolution of Bayesian Posterior Distribution for a Simulated Inspection.
Results after each group of 5 trials are plotted.

The actual simulation depicted resulted in 4 misses in the 45 trials for an average probability $p_m = 0.911$. The evolution of the posterior distribution shows that it is quite broad after the initial sets of trials, but rapidly becomes more peaked around the mean probability value. The confidence level for any value of the probability of detection p can be obtained directly from the posterior distribution. For comparison, the evolution of the estimates for the mean and 95% p_a are shown in Figure H-2. For such a small number of trials, the 95% p_a is well below 0.90.

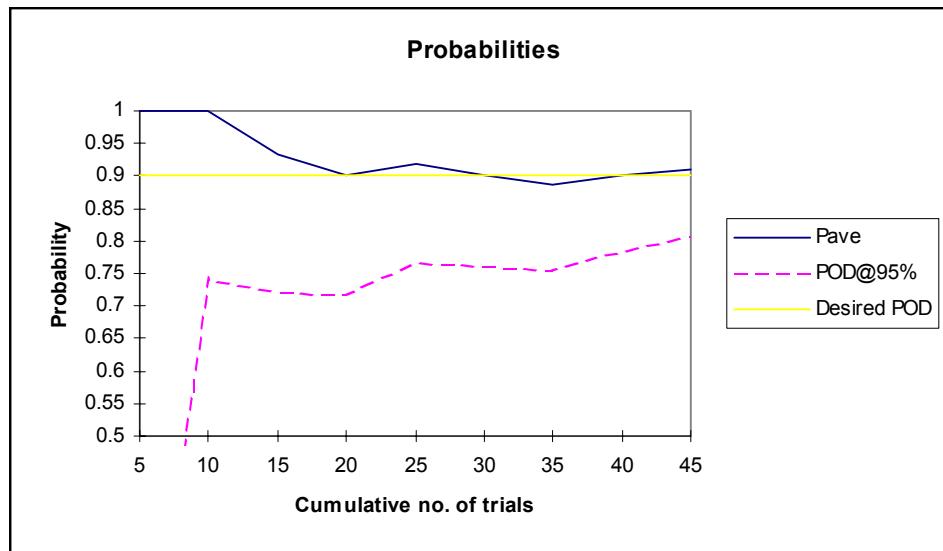


Figure H-2: Binomial POD at 95% Confidence for the Simulation shown in Figure H-1.

The three safety levels are shown in Figure H-3. It can be seen that the true safety level does indeed attain the desired 0.001. The Bayesian estimate of the safety level is conservative, however, it is significantly closer to the real value than the classical estimate from the 95% lower bound on the POD.

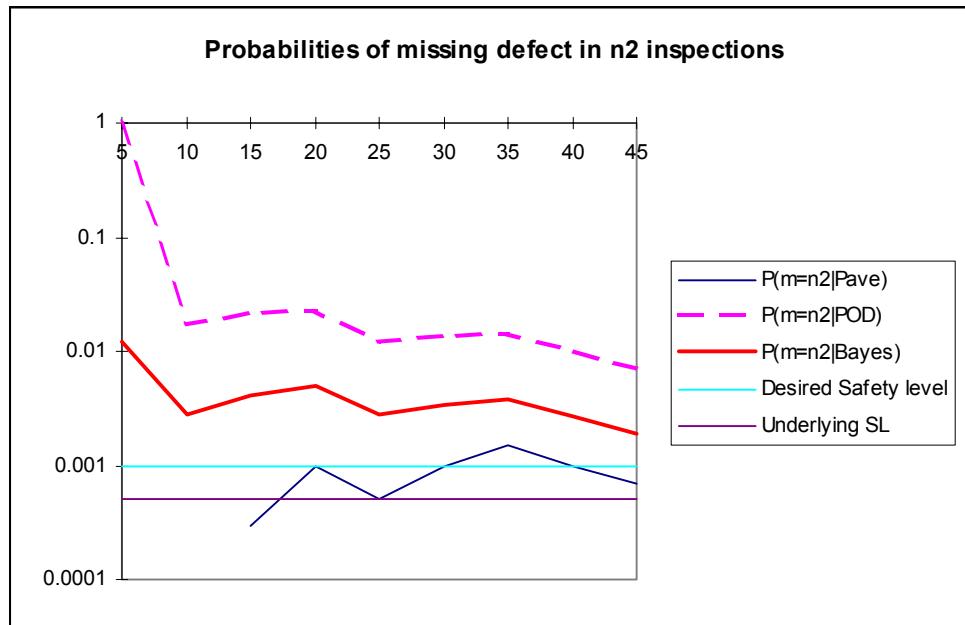


Figure H-3: Safety Levels for the 45 Defect Simulation.

The greater efficiency in translating the full available information on NDT reliability into a direct estimate of the safety level which can be expected, offers the possibility that useful reliability statistics and safety level estimates can be generated from substantially less data than would be required for the standard POD analysis. This approach requires further investigation.

The binomial and Bayesian methods were also applied to the F-16 data supplied by the RNLAf. Initially, this was supplied as a simple record of hits and misses without crack size information. The results are summarised in Table H-2.

Table H-2: Analysis of RNLAf Data

ASIP #	A/M	h	m	Safety			Bayesian estimates			Ratio
				Pav	POD95%	level	POD95%	Pmean	Safety level	
1001	M	3	0	1	0.362	0.260	0.47	0.84	0.029	0.11
1004	M	5	0	1	0.54	0.097	0.61	0.89	0.012	0.12
3005	M	39	51	0.43	0.344	0.282	0.35	0.43	0.185	0.66
4004	M/A	33	13	0.72	0.588	0.070	0.60	0.71	0.029	0.41
8032	M	37	0	1	0.922	0.0005	0.92	0.98	0.0001	0.20
8033	M	9	0	1	0.714	0.023	0.74	0.93	0.003	0.15
8104	M/A	1	0	1	0.05	0.857	0.22	0.71	0.100	0.12
8106	M	3	8	0.27	0.079	0.781	0.12	0.30	0.363	0.46
8107	M	3	1	0.75	0.249	0.424	0.34	0.69	0.071	0.17
8108	A/M	1	0	1	0.05	0.857	0.22	0.71	0.100	0.12

The Bayesian “POD” is identified as the probability corresponding to the lowest 5% of the posterior distribution. The safety level is, in each case, the probability of making three misses in three inspections. It can be seen that while the Bayesian and binomial PODs are similar, the safety level predictions are distinctly different. The final column labelled SL Ratio is the Bayesian safety level divided by its binomial counterpart. For the largest group of cracks, 90 inspections of 39 defects, the binomial SL is only 50% higher than the Bayesian, although neither is close to the desired 0.001 level. In the other groups the ratio is considerably larger.

H.1.3.1 Extension to Size-Dependent Probabilities of Detection

The discussion of binomial and Bayesian methods in the text above assumed that there was a single probability of detection for the defects. In practice, of course it is assumed that the probability will depend on the defect size and possibly other defect features.

The most straightforward method of introducing consideration of crack size effects is to introduce a threshold crack length and then to deal with only those inspections which yielded “hits” and “misses” of cracks greater than this length. If several thresholds are used, the cracks can be binned into discrete ranges. Using all of the cracks above, each threshold is referred to below for convenience as the cumulative method.

The probabilities calculated from all observations of cracks above a threshold size is not a straightforward probability function of crack size like the POD(a) assumed in the main report, Section 5.2. It clearly depends implicitly on the population of cracks used in its determination. In the case of in-service defects, the defect population at any time is determined by the aircraft’s operational usage and therefore represents the crack population which it is necessary to detect. As the defects can be expected to start small and grow progressively larger, the crack population size distribution for the cracks above any threshold should approach a stationary distribution where the detection rates balance the new crack growth. The errors associated with the use of the data at short times will be conservative due to the presence of increased numbers of small cracks. The probability calculated will be the best estimates for detection of any of the population of defects above the threshold size chosen at random. This is not quite the same as the probability of a single defect being detected in successive inspections along a deterministic growth trajectory, but given the limited data available, it is likely to be as good as, if not better, than any other measure.

The alternative method of performing all of the calculations independently for each of the crack size intervals is known as the “range interval method”. The original methods for determining PODs proposed by the American Society for Non-destructive Testing (ASNT) suggested that the range interval method should be used. This was usually found to be too inefficient (or expensive) as it requires large numbers of specimens in each crack range to give high POD estimates. It also gives results which are dependent on the selection of the intervals. It was therefore replaced by the curve fitting methods described in the main report, Section 5.2, which have become standard. For illustrative purposes, the range interval method calculations are also shown below.

H.1.4 Examples/Case Study

The data used for these examples is that provided by the RNLAF F-16s for ASIP station 3005, which is extensively considered elsewhere in this report. The data provided included three sets:

- a) 39 hits and 51 confirmed misses calculated using the assumed average usage spectrum
- b) 39 hits and 51 confirmed misses calculated using individual aircraft usage spectra

c) 39 hits and 215 unconfirmed misses, the maximum number which could have occurred, calculated using individual aircraft usage spectra

Initial analysis of data set a) was carried out using a curve fitting approach with a log-logistic POD curve. This showed that the mean curve reached 90% at a value of $a_{90/50} = 0.090$ inch, while the 95% confidence curve only just reached 90% within the range of the defect sizes at $a_{90/95} = 0.203$ inch.

To test the performance of the Binomial and Bayesian methods using variable thresholds, the data were binned into 11 discrete size ranges. The lowest range was 0.00 to 0.02 inch. The remaining thresholds were set at intervals of 0.01 inch. The distributions of inspections, misses and hits in each bin is shown in Figure H-4, where bin 1 corresponds to the largest cracks and 11 to the smallest.

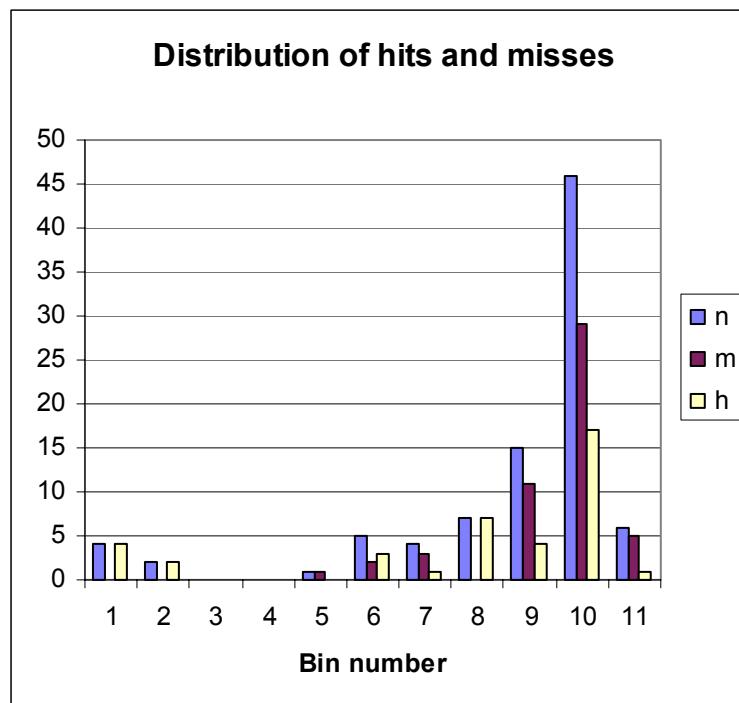


Figure H-4: Distribution of Inspections, Misses and Hits for F-16 3005 Data Set a).

It can be seen quite clearly that there are few large cracks. The longest crack which was actually missed had a length of 2.0 mm (0.08 inch).

H.1.4.1 Data Set a) Average Use Spectrum

The results of applying the cumulative method to the data of set a) are summarised in Figure H-5 and Figure H-6.

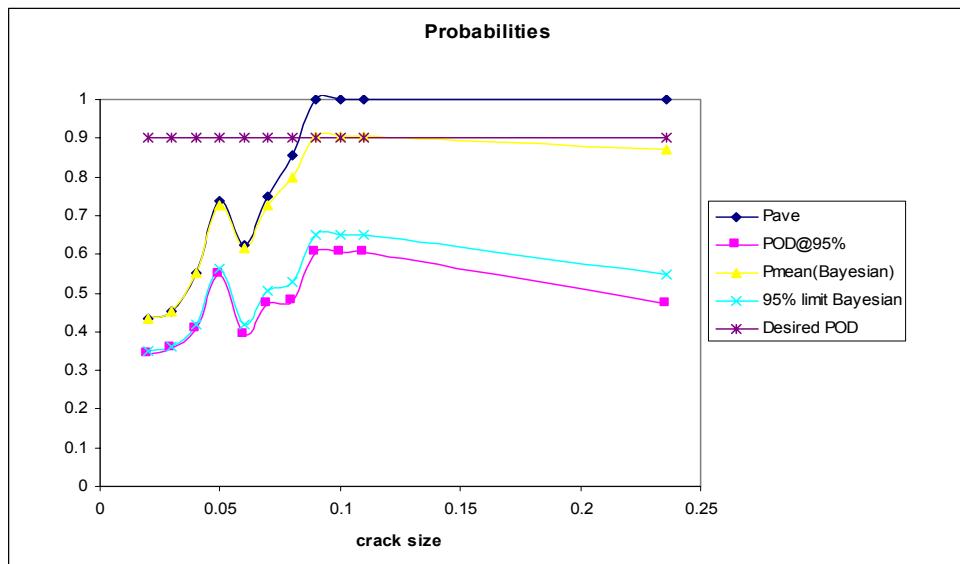


Figure H-5: Probabilities of Detection for Cracks above a Crack Size Threshold.

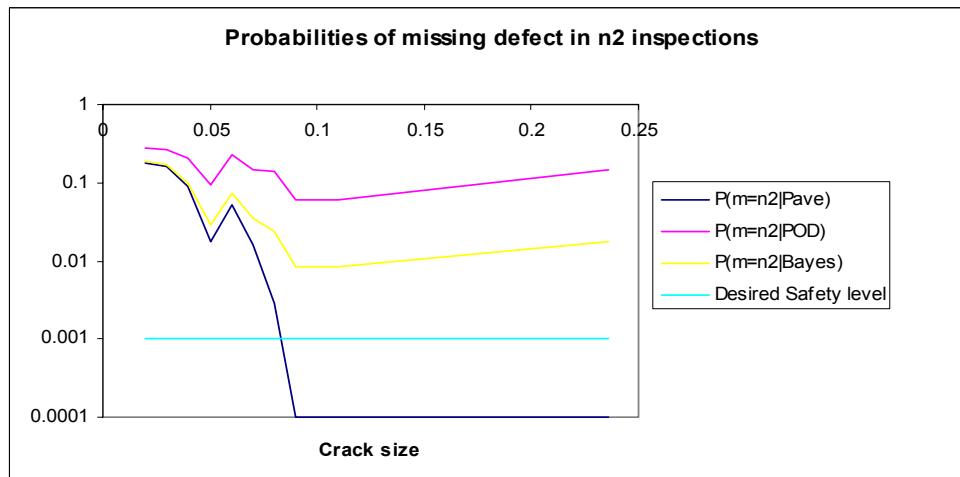


Figure H-6: Probability of Missing Defect Completely in Three Inspections for Cracks above the Crack Size Threshold.

It can be seen that the estimates for the mean and 95% confidence limits converge for the large numbers of inspections (90) at the small crack thresholds. This shows that the improvement of up to an order of magnitude in the estimated safety level arises from the interpretation of the data directly to predict the outcome of the three inspections and not from an artificial overestimate of the capabilities of the technique by the Bayesian method.

The estimated safety levels are lower than the desired 0.001 possibility of failure. Taking a threshold at 0.09 or 0.10 inch, the latter being the desired critical crack size for detection, the figures show that the inspection interval would have to be halved to allow six inspections during the growth time to ensure the

0.001 safety level is obtained from the inspections. As more data is obtained, either from more experience in operating the F-16 in The Netherlands, or from pooling experience with other air forces, this restriction would be expected to be relaxed.

Alternatively, building in prior knowledge about the inspection task, the safety level estimate can be improved to the extent that three inspections may be shown to suffice. This prior knowledge could be the result of initial trials or previous experience with similar inspections, however this approach of using prior knowledge requires further justification.

The probabilities of detecting defects above a threshold defect size were compared to the POD curves generated by the standard methods fitting a log-logistic function using the maximum likelihood method. The appropriate curves are shown in Figure H-7.

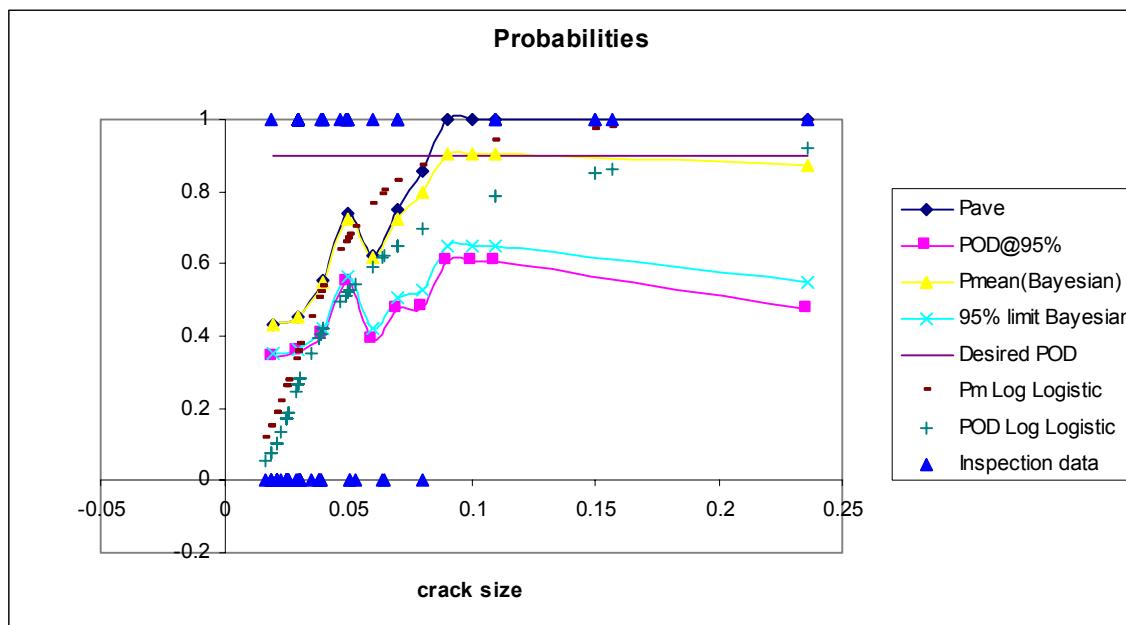


Figure H-7: Probabilities of Detecting Cracks above a Threshold Crack Size Compared to Log-Logistic POD Curves.

Three observations can be made from this comparison:

- 1) The standard method only just achieves a 90/95% POD within the range of crack sizes. The $a_{90/95}$ value obtained, 0.203 inch, is more than twice the desired value of 0.1 inch, even though the fitted “mean” curve crosses the 90% level at 0.09 inch.
- 2) The extreme point on the 95% confidence limit curve must be subject to fluctuations, therefore we could not expect to rely on the standard method achieving a 90/95% level within the crack size range (See discussion of b and c data sets below).
- 3) Many of the detected cracks are small. Over half of the cracks are detected at a size of no more than 0.04 inch and three quarters are detected by the time they have reached 0.06 inch. At these crack sizes, the mean probability curve has the values 0.50 and 0.79, respectively, while the 95% curve has values 0.39 and 0.62. This shows that the a_{90} values are being estimated in the regime of primarily short

ANNEX H – BINOMIAL AND BAYESIAN ANALYSIS

crack data. As assessed by the method in the main report, Section 5.2, this makes the estimates particularly susceptible to statistical fluctuations, and the value of the $a_{90/95}$ should be increased if a true 95% confidence is to be maintained.

It was noted earlier that while the Bayesian method is an elegant way of predicting the probability of missing a defect in a set of in-service inspections, it is certainly not the only way. The simplest is to note that the probability of outcome of an in-service inspection based on previous results from trials or in-service experience can be obtained from a contingency table. Using the normal approximation with a simple correction for the small numbers involved (Yates' correction), a third estimate of the safety level can be obtained from the χ^2 -squared distribution. This is shown in Figure H-8. As can be seen from the figure, the χ^2 estimate lies between the binomial POD and the Bayesian estimate. (Without the correction the χ^2 estimate is non-conservative and leads to safety levels even higher than the Bayesian.)

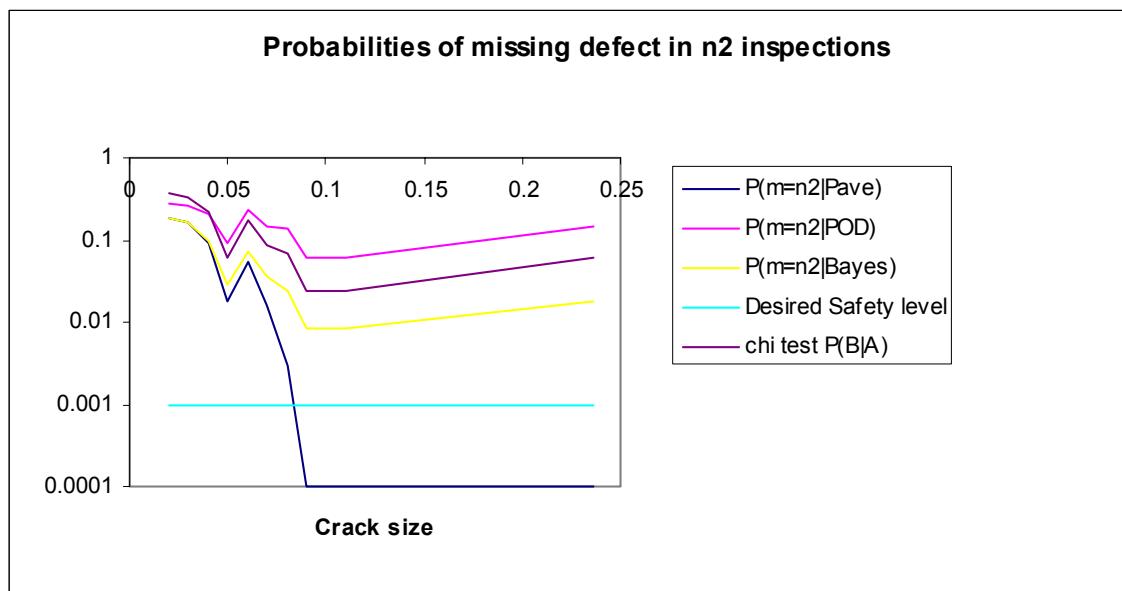


Figure H-8: Safety Level Estimates for Three Inspections including the Level Calculated from the Chi-Squared Test.

For completeness, the range interval method was considered briefly. It is immediately apparent that there is insufficient data to use this approach for this inspection. Nevertheless, again using 0.01 inch bins, the probabilities and corresponding safety levels shown in Figure H-9 and H-10 were obtained.

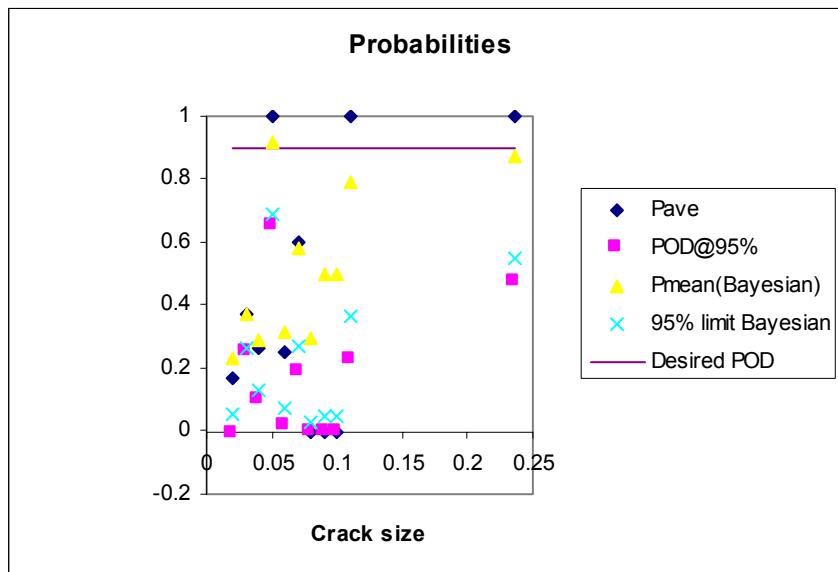


Figure H-9: Binomial and Bayesian Probability Estimates from the Range Interval Method.

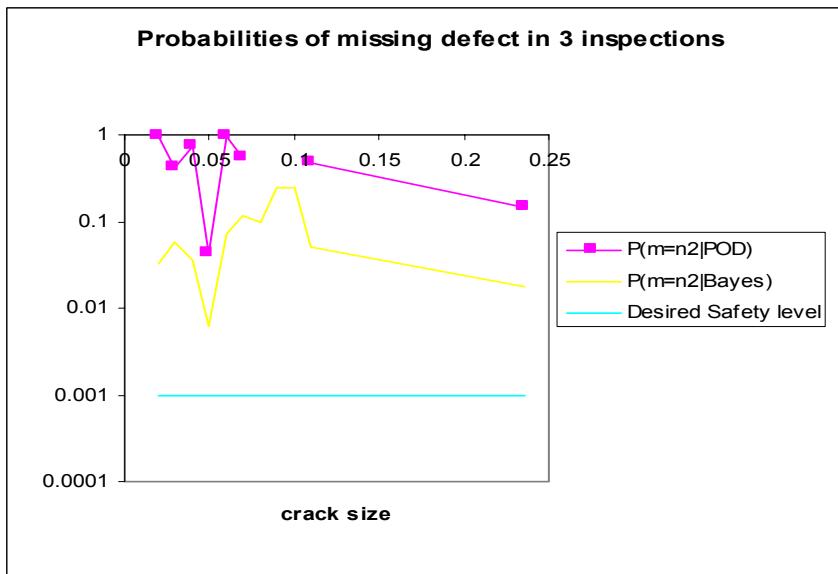


Figure H-10: Safety Level Estimates from the Range Interval Method.

It can be seen that substantially more data would have to be available for this approach to be useful.

H.1.4.2 Data Set b) Individual Aircraft Use Spectrum

In an attempt to improve the estimates of missed cracks, the back-projection calculations for previous inspection times were repeated using data relating to the load spectra experienced by individual aircraft, labelled SCSI. This second data set therefore consists of the same 39 hits at the measured crack length and the

ANNEX H – BINOMIAL AND BAYESIAN ANALYSIS

same 51 misses, but with their lengths recalculated. In practice, this makes little difference to the distribution of misses, see Figure H-11.

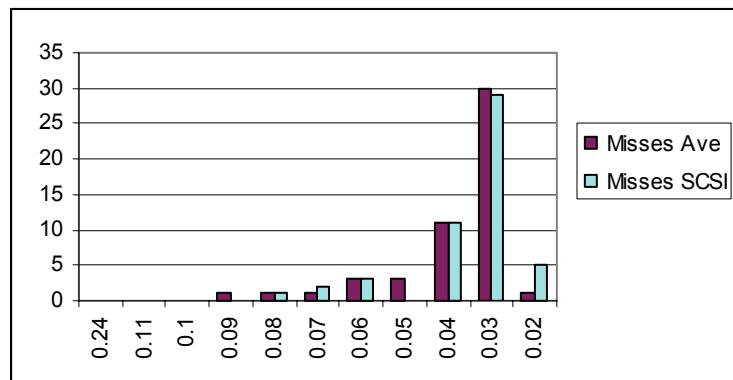


Figure H-11: Distribution of Misses in the Average use and SCSI Data Sets.

The differences in missed crack lengths may appear slight, but it was sufficient to alter the a_{90} crack length estimates significantly. The new value for the mean a_{90} obtained from the log-logistic model is $a_{90/50} = 0.133$ inch and $a_{90/95} = 0.503$ inch, well beyond the range of crack size data.

The changes in the missed crack sizes had almost no effect on the cumulative or range interval methods. The probabilities and safety levels obtained are shown in the following figures (Figure H-12 through H-15).

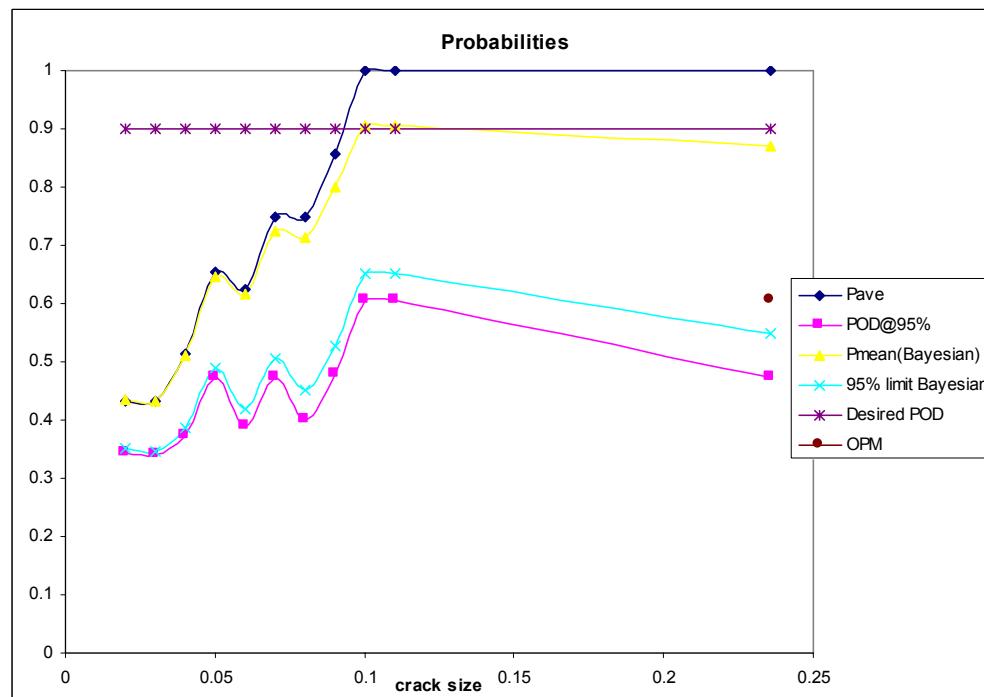


Figure H-12: Probabilities of Detection for Cracks above a Crack Size Threshold from the b) Data Set.

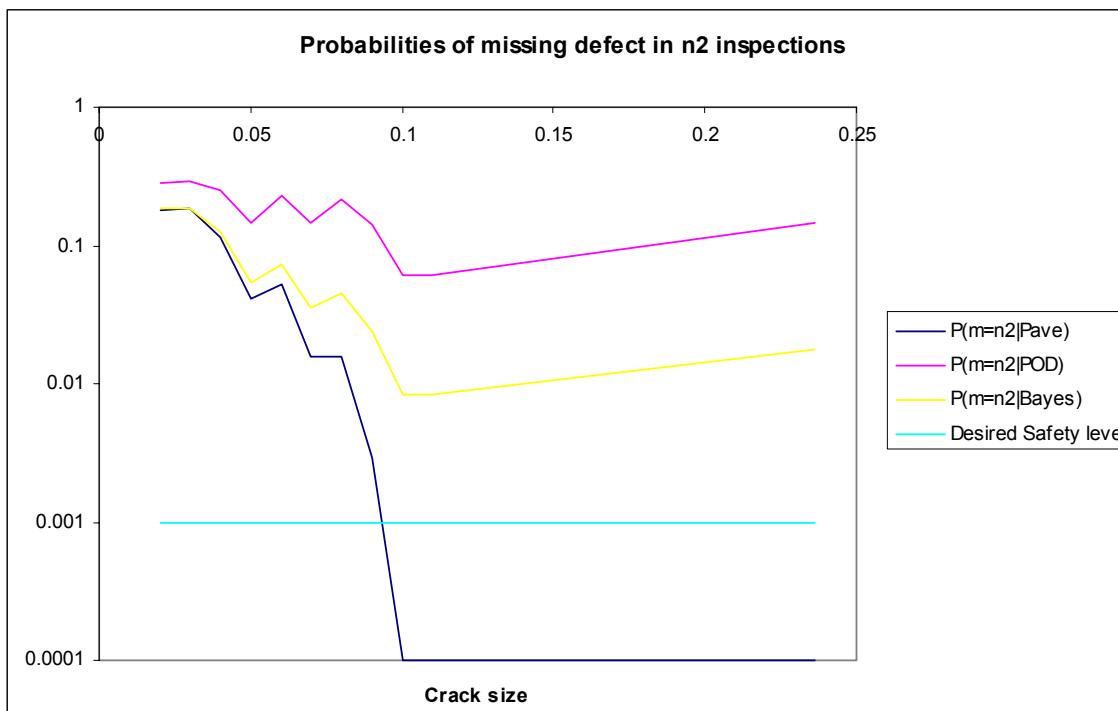


Figure H-13: Probability of Missing Defect in Three Inspections for Cracks above a Size Threshold for the b) Data Set.

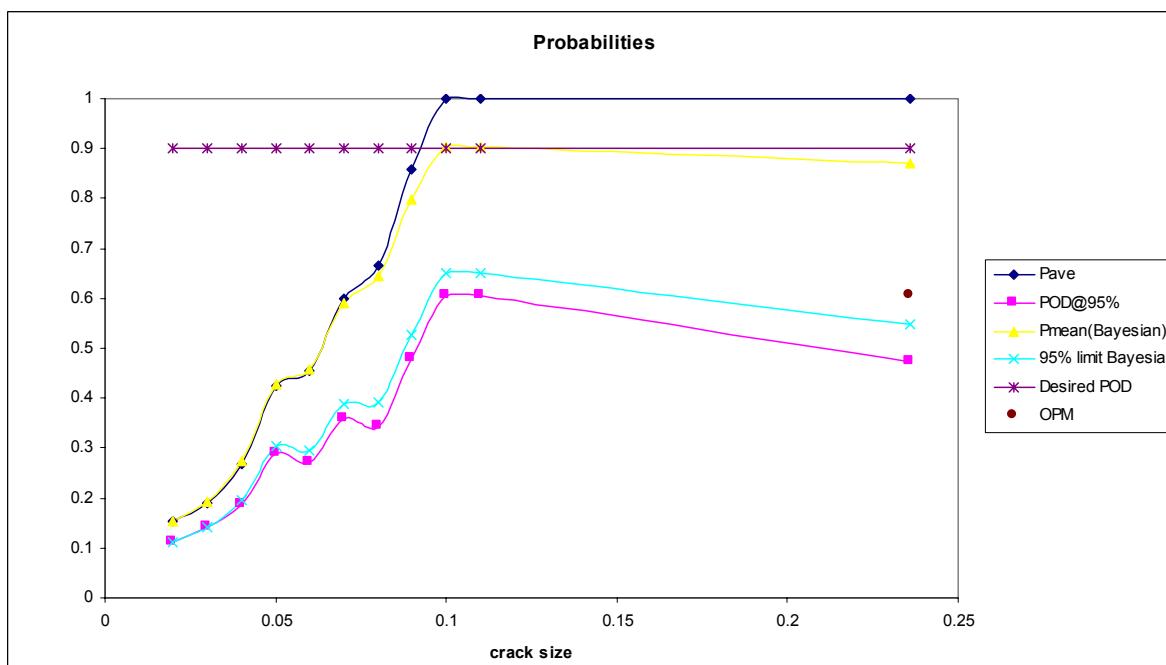


Figure H-14: Probabilities of Detection for Cracks above a Crack Size Threshold from the c) Data Set Incorporating all 215 Potential Misses.

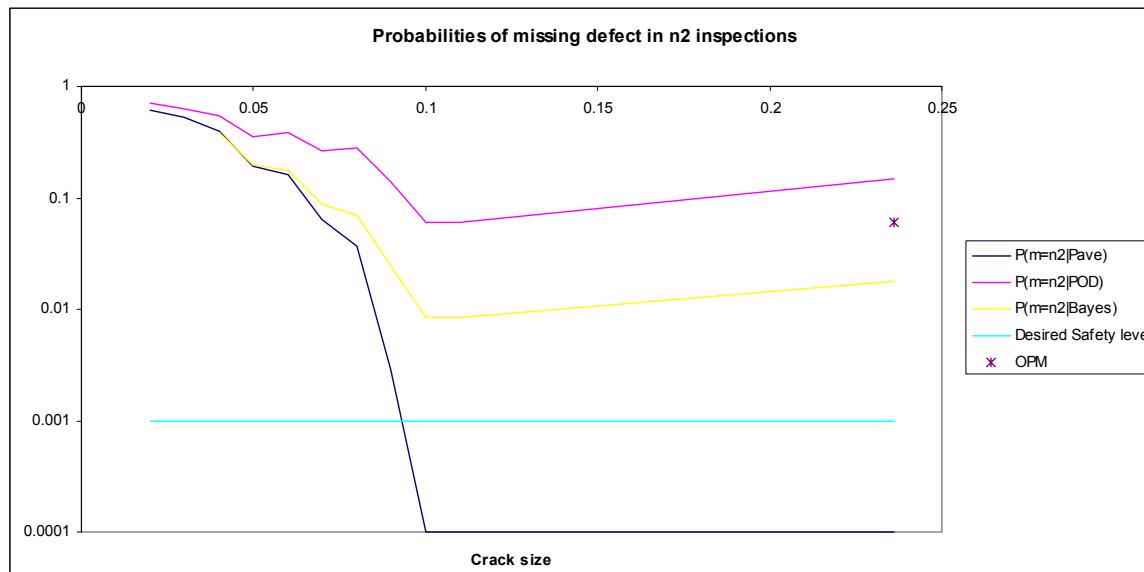


Figure H-15: Probability of Missing Defect in Three Inspections for Cracks above a Size Threshold for the c) Data Set.

H.2 SUMMARY

More efficient statistical methods can demonstrate higher safety levels than the standard analysis. This may not be necessary for situations where there is adequate reliability data to use the standard methods. It may be crucial to reliance on NDT where verification of high reliability is limited by available data.

One approach based on Bayesian inference has been shown to be able to give useful quantitative estimates for safety levels on very limited data. Further analysis of this, or other approaches which make the best use of limited data, should be undertaken to provide a more flexible alternative to the standard methodology.

The Bayesian approach in particular, through the estimation of the posterior distribution, gives the maximum information on the inspection and may be used as a check on the effectiveness of the technique in service.

Annex I – SIMULATIONS OF THE EFFECT OF MISSING MISSES IN POD ESTIMATION FROM IN-SERVICE DATA

I.1 INTRODUCTION

It has been recognized for many years that application parameters and human factors, as described in the Reliability Formula proposed at the First European-American Workshop on NDE Reliability, are limiting factors in the performance of non-destructive evaluation (NDE). Experiments to determine the probability of detection (POD) for specific NDE applications are often designed and performed in conditions that do not reflect the conditions in which the actual inspections are performed. Climate, training and motivation are just a few examples of variables that are difficult to mimic in experiment.

In 1981, David Simpson of the National Research Council Canada proposed the use of “field” inspection data for accurate determination of POD [1]. Data collected during normal maintenance actions offers potentially reduced cost of collection and an accurate reflection of application parameters and human factors. However, there are new issues raised by this process including, for example: crack size determination; missed cracks in service; and statistics of small data sets that are not normally present in laboratory experiments. These factors and others affect the confidence in the calculated POD, and must be quantified before POD data of this type can be used (see for example Leemans [2], Bruce [3] and Forsyth et. al [4]). These difficulties have prevented the wide-spread use of field data for POD estimation, however, a few studies have incorporated elements of this methodology [5,6].

Recently, work by Spencer [7] showed that “hit/miss” type of NDI data collected from the field will always yield non-conservative estimates of the POD of the system which was used to generate this data. In order to assess the practical implications of Spencer’s result, this report uses lifting data from actual aircraft situations to:

- 1) evaluate the level of non-conservative bias in field estimation of POD from “hit/miss” type NDI data, and
- 2) evaluate field estimation of POD using “ \hat{a} vs. a ” type NDI data.

It is interesting to note that this possibility is raised by Goranson [8] in reference to how Boeing develops the “Damage Tolerance Rating” used in the maintenance of their commercial fleet. No analysis of the effect of this potential bias is provided.

I.2 THEORY

I.2.1 Probability of Detection

The most common method for quantifying the reliability and sensitivity of an NDT system is probability of detection (POD) analysis. In brief, POD analysis provides a methodology for estimating the detection capability of an inspection method as a function of crack size. Following the analysis of Berens [9], the POD at a crack of characteristic size “ a ” is defined to be the average probability of detection of all cracks at the size “ a ”. This definition reflects the fact that the detectability of cracks will vary with a number of factors, including but not limited to size. Therefore, the POD curve is drawn through the mean POD for each crack size, and the confidence level associated with the POD curve reflects the fact that the curve was calculated using a sample population.

Numerous statistical methods have been proposed to estimate this relationship for two different types of NDT data: “hit/miss” type of data and “ \hat{a} vs. a ” type of data. The “ \hat{a} vs. a ” type of data refers to an NDT system that provides an estimate \hat{a} of the crack size a , when a crack is found during an inspection. The “hit/miss” type of data refers to an NDT system which gives results as either indicating the presence of a crack (a hit) or the lack of a crack (a miss) on the inspection subject. Most NDT systems provide some estimate of crack size, however, in field applications, the inspection data is usually recorded as “hit/miss”.

I.2.2 Estimation of Probability of Detection

Two general types of statistics for the estimation of the POD-crack size relationship have been proposed in the literature. The two categories are binomial methods (e.g. [10]) and curve-fitting methods (e.g. [11]). Neither of these methods take false call rates into account. Both methods are still in common use. In this work, only the curve-fitting methods are used, specifically those detailed in the United States Department of Defense MIL-HDBK-1823 (1999 revision).

For “ \hat{a} vs. a ” type data, it has been noted in many cases that the logarithms of \hat{a} and a are linearly related, with residuals normally distributed with mean zero and standard deviation δ^2 , and one can write:

$$\hat{\ln a} = \beta_0 + \beta_1 \ln a + \varepsilon \quad (\text{I-1})$$

and therefore

$$POD(a) = 1 - Q\left[\frac{\ln a - \mu}{\sigma}\right] \quad \text{where } \mu = \frac{\ln(y_{th}) - \beta_0}{\beta_1} \text{ and } \sigma = \frac{\delta}{\beta_1} \quad (\text{I-2})$$

Recently, Spencer [12] proposed an extension to the curve-fitting method of estimating POD, which included intrinsic minimum and maximum POD values, based on false call rates and false miss rates. This model can be written as

$$POD(a) = p_h + (1 - (p_m + p_h)) \cdot F(a; \mu, \sigma) \quad (\text{I-3})$$

where $POD(a)$ is the probability of detection at the crack size a , p_h is the false call probability, p_m is the probability of missing a crack independent of crack size, and $F(a; \mu, \sigma)$ is the two-parameter distribution used to fit the data from equation (I-2). This distribution is usually a two parameter (μ, σ) log-normal curve as in equation (I-2) – log-logistic curves are also used.

I.2.3 The Use of Field Data for POD Estimation

Field inspection data can be employed to obtain “hit/miss” data which can be used in turn to estimate the POD for that particular inspection. Inspections at a particular site are recorded over time or operational cycles, until a crack is found. This gives a hit, and an estimate of the crack size is usually obtained either from the inspection result itself or by performing a secondary inspection, or by disassembly and verification tests.

The preceding inspections at this site can be used to predict miss points, by estimating the size of the crack at the inspections performed before the crack was found. This is a complex procedure that requires knowledge of initial crack sizes, crack growth rate and the loading on the site.

Five main problems have been identified with the use of field data for POD estimation. These are:

- 1) uncertainty in “back-casting” crack sizes
- 2) uncertainty in crack growth
- 3) crack size estimation at time of detection
- 4) uncertainty in operational conditions
- 5) POD model sensitivity to small sample sizes

Various authors have examined the above problems (see for example [2,3,4]). This report examines the additional complication of the inherent non-conservative bias in POD estimates from field data of “hit/miss”-type, and tries to determine if any such bias exists when using signal-response type data.

I.3 EXPERIMENTS

Monte Carlo methods were used to simulate the lives of sets of components. The simulations were based on a fifth-stage compressor disk from the J85-CAN40 engine. In the late 1980’s, the Institute for Aerospace Research (IAR) was involved in a program sponsored by Canada’s Department of National Defence to convert the maintenance of these components from safe-life to retirement for cause. In the course of this program, crack populations, crack growth and reliability of available NDI were all determined. Both deterministic and probabilistic fracture mechanics were applied (see Koul et al. [13] for further details). Only the probabilistic fracture mechanics results are considered here.

The life limiting element of the J85-CAN40 fifth-stage compressor disk is low cycle fatigue cracks in the bolt holes. The starting point for the simulations in the modelling of these components was a “time to crack initiation” (TTCI), following the original approach for the engine maintenance. The manufacturer provided an experimentally-derived distribution of the time to an “initial crack size” of 0.8 mm for the bolt holes under the expected loading conditions. The relationship between the cyclic stress intensity factor ΔK and the crack size a was determined through three-dimensional finite element analysis [14]. Crack growth data for the bolt holes in the components under consideration was also determined experimentally and fit to a modified Paris Law expression, details are also available in [13].

The critical or dysfunction crack size was 4.27 mm. The distributions of times to the “initial” and critical crack sizes are shown in Figure I-1. It is interesting to note that there is significant overlap in these distributions, which is an indication that the safe-life approach is not going to be very efficient – many components will have small or no cracks at times when a small number of components are at the end of their life.

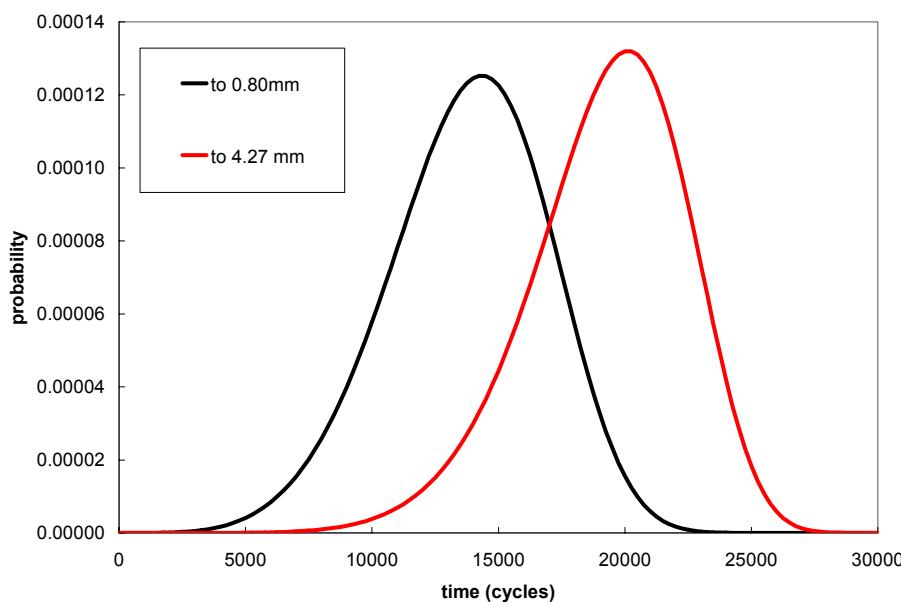


Figure I-1: The Distributions of the Time (in cycles) to the “Initial” Crack Size of 0.80 mm and the Dysfunction Crack Size of 4.27 mm.

The results are presented in the following sections. In each chart, the “underlying” mean POD used for the simulation is shown by the solid red line and the POD which would be estimated from the available field data is shown in green. If somehow the undetected cracked components were available for analysis, the POD estimate would be that shown by the blue line. The known crack data is shown as green points, and again, the actual population which is not known in practice is shown as blue points.

I.3.1 “Hit/Miss” Data Simulations

Sets of component life simulations were performed using two different simulated inspection techniques, in order to determine if the underlying POD affects the non-conservative bias found in field estimation of POD from “hit/miss” type data. The first inspection technique represents a very good technique in terms of high signal-to-noise ratio (SNR) at the level of the detection threshold. In terms of the POD relationship, this results in a very steep curve. The second inspection technique represents a lower SNR at the level of the detection threshold, more representative of highly manual and operator-dependent NDI methods.

For each simulated inspection technique, the lives of a set of 50 components are simulated from zero cycles until a crack is found during a scheduled inspection. Inspections are simulated every 1000 cycles, starting at 2000 cycles for the “good” inspection technique, and 5000 cycles for the worse technique. Finding a crack means the component is retired. The POD of the inspection is estimated at each inspection interval from the data which would be available at that time. For example, if at 5000 cycles, 25 components of the 50 have been found to have cracks, there will be 25 hits and all the previous misses available for POD estimation.

I.3.1.1 “Hit/Miss” Inspection Technique 1 – High SNR

The evolution of the in-service inspection data and estimated POD are shown in Figure I-2 through to Figure I-5. It can be seen in this case that the estimate of POD from the field data would be reasonable, if slightly

non-conservative. Note that no data is available until the fleet has reached 6000 cycles of usage. As this fleet continues to operate, the POD estimates remain very close to the underlying POD. When all the fleet has been retired, the estimated POD from the field is shown in Figure I-5 as the green line, the underlying POD shown in red.

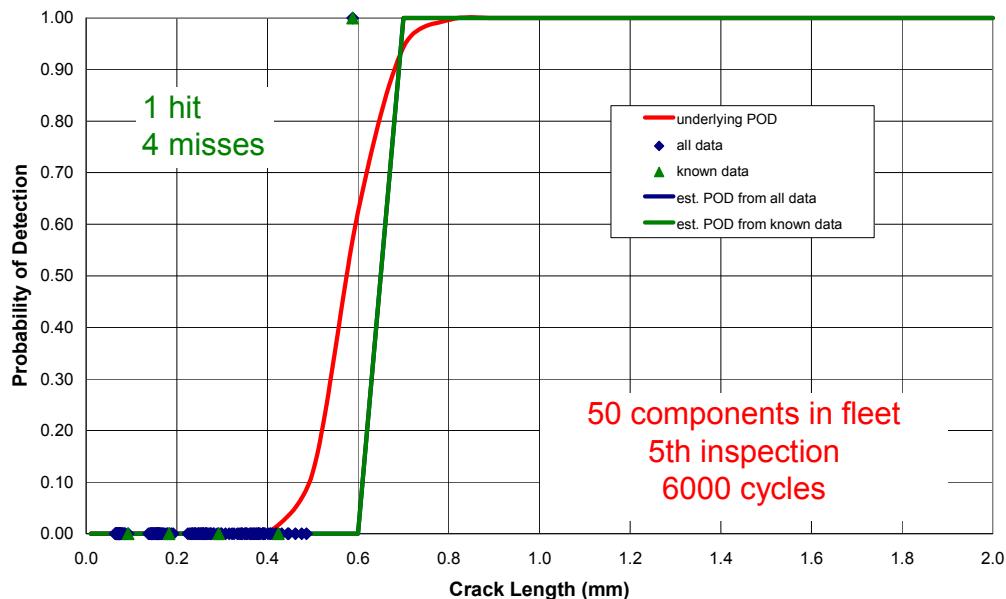


Figure I-2: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

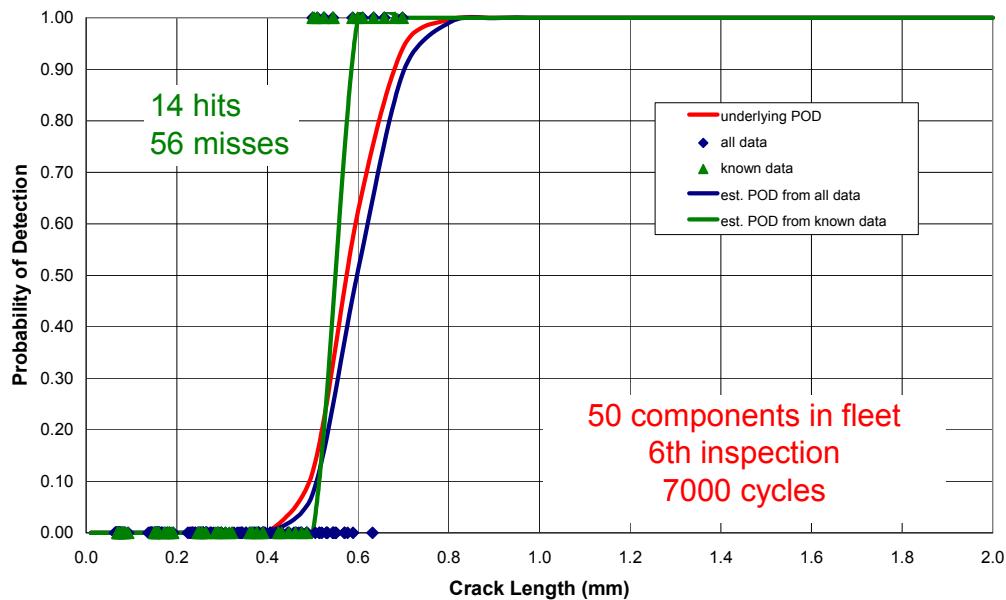


Figure I-3: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

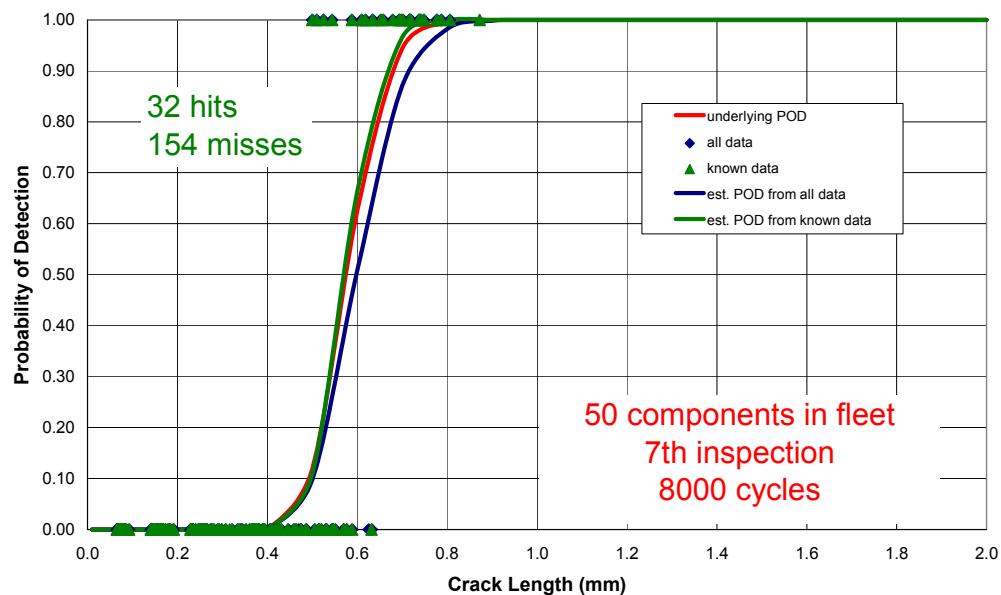


Figure I-4: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

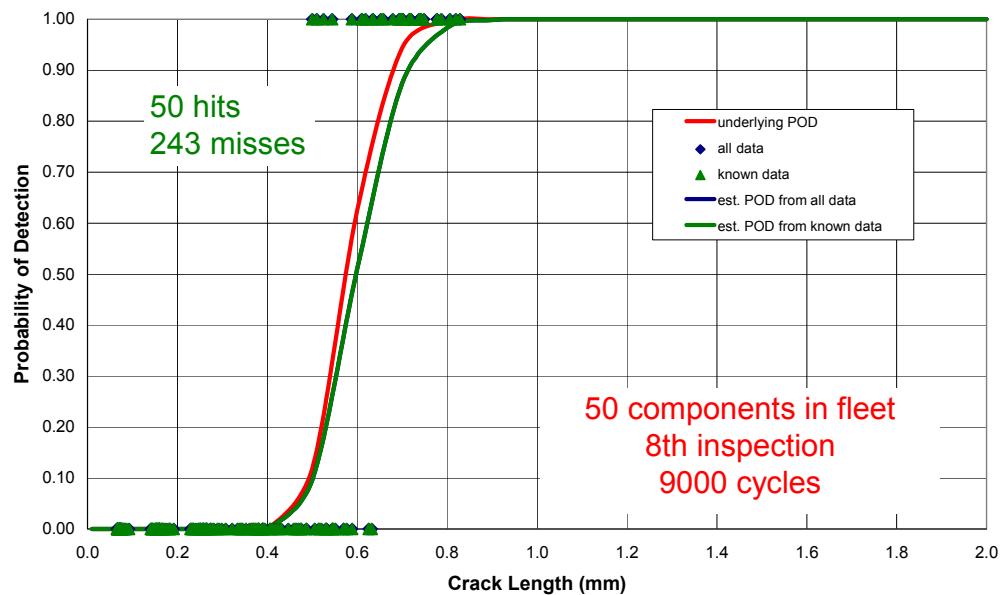


Figure I-5: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

I.3.1.2 “Hit/Miss” Inspection Technique 2 – Low SNR

The evolution of the in-service inspection data and estimated POD are shown in Figure I-6 through to Figure I-11. At the early times, the POD estimate is very non-conservative. In fact, for this simulation, the estimated POD only begins to approach the underlying POD once nearly all the components in the fleet have been retired, as shown in Figure I-10 below.

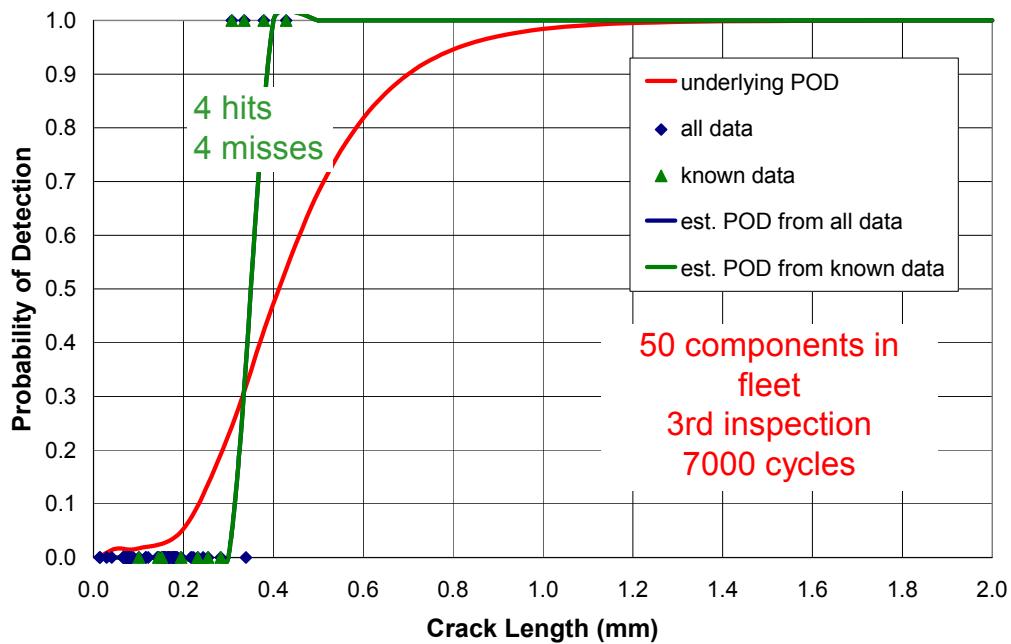


Figure I-6: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

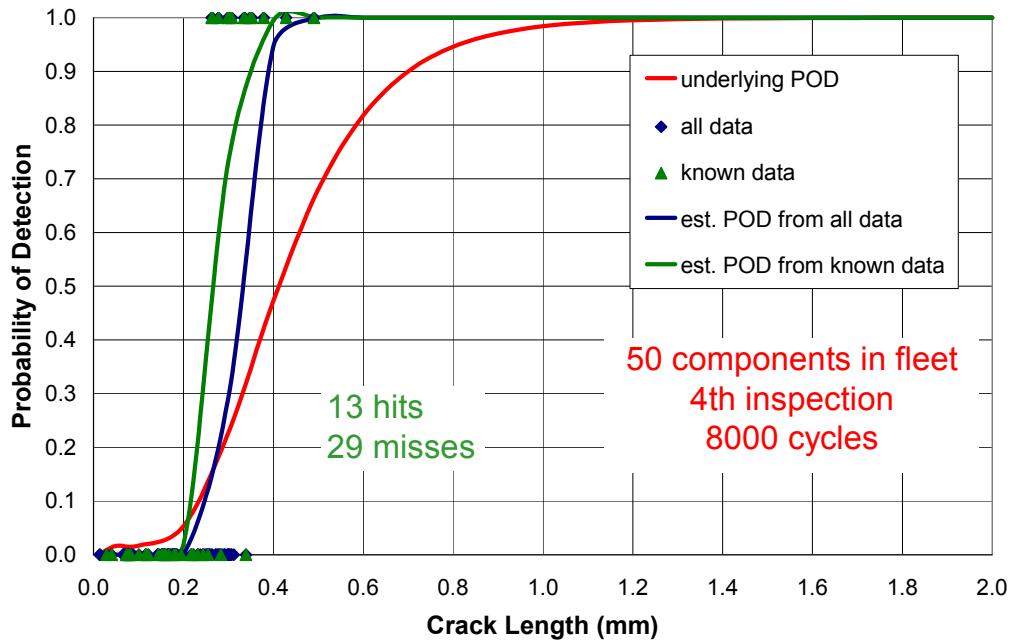


Figure I-7: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

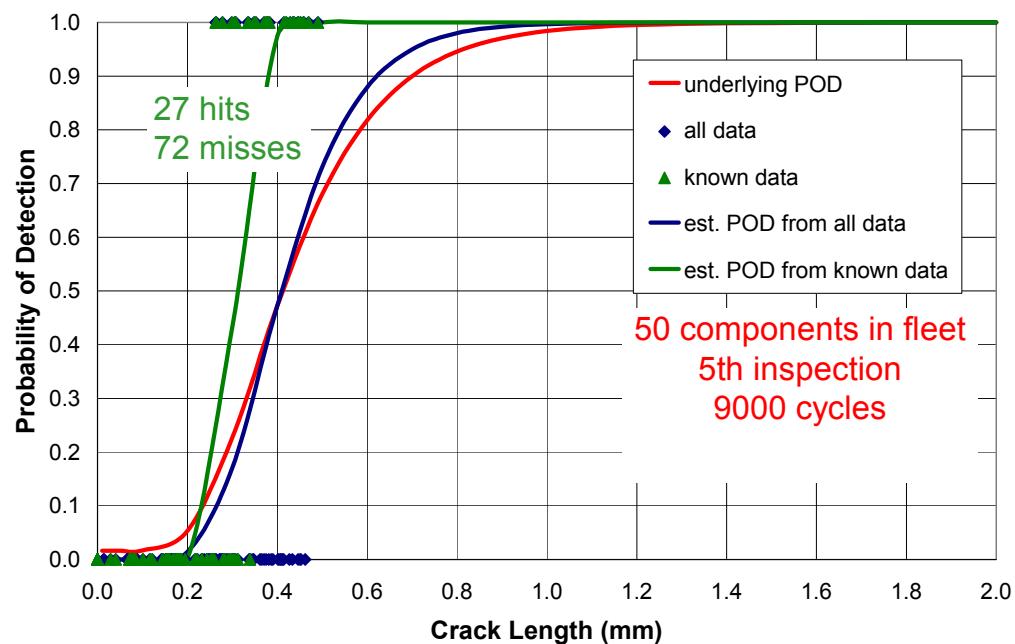


Figure I-8: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

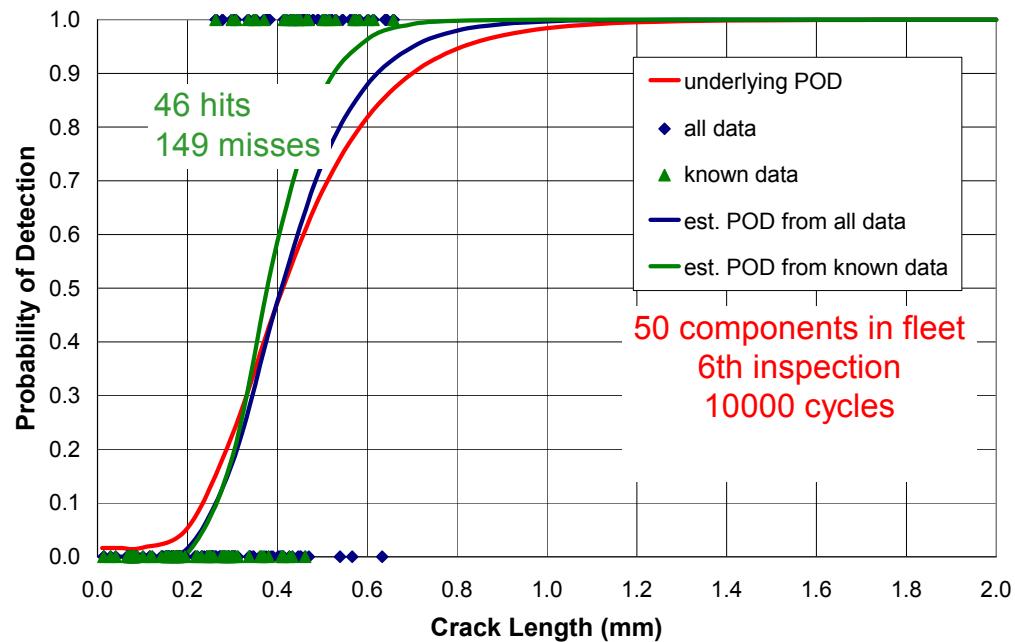


Figure I-9: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

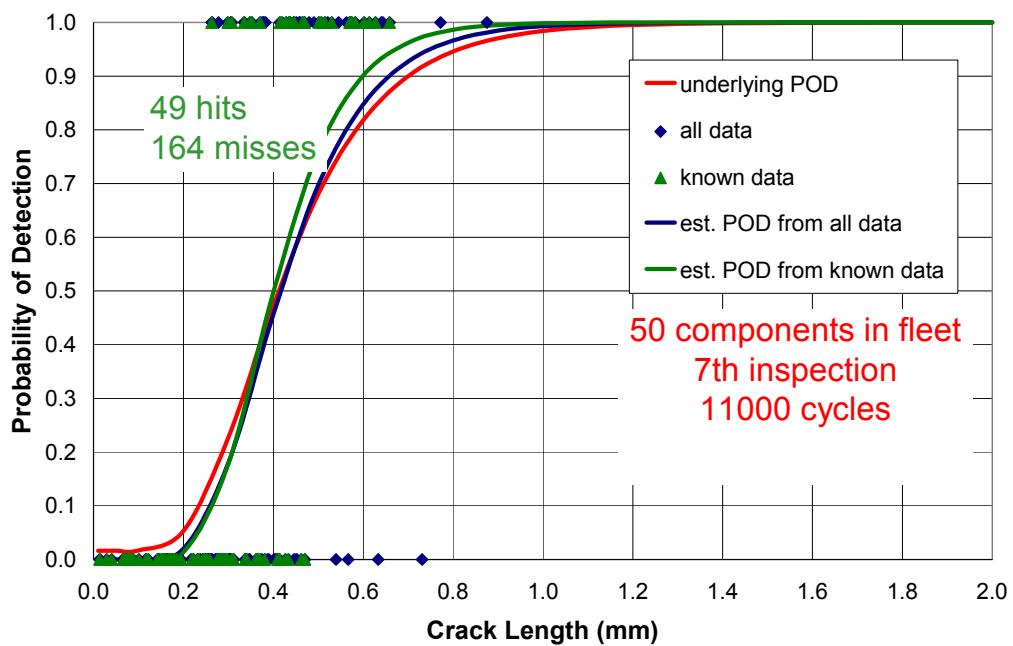


Figure I-10: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

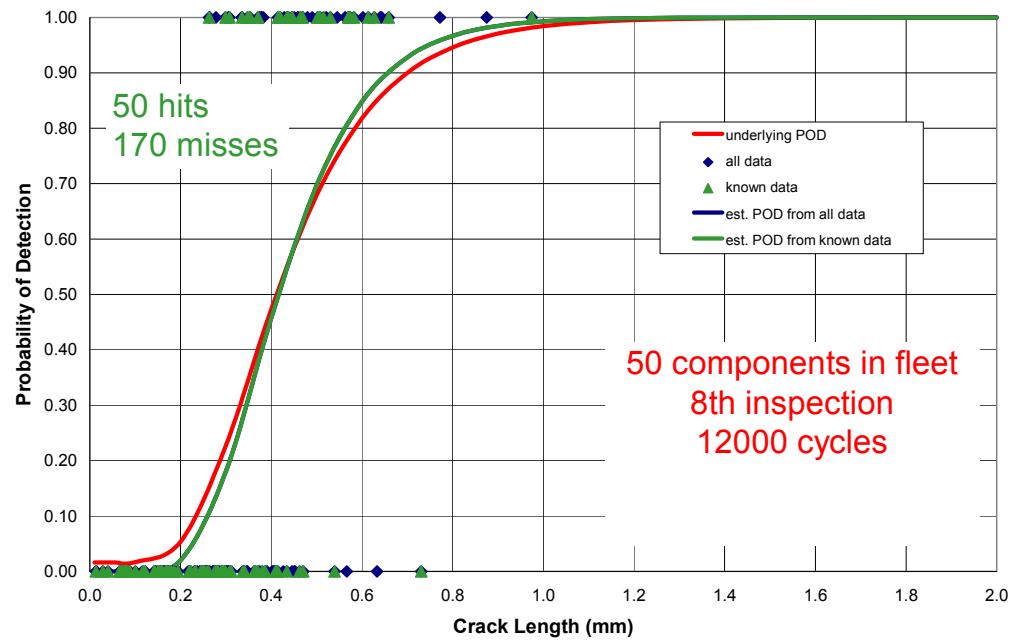


Figure I-11: “Hit/Miss” Data and Estimated Mean PODs Compared with the Underlying POD.

I.3.2 “ \hat{a} vs. a ” Data Simulations

Sets of component life simulations were performed using two different simulated inspection techniques, in order to determine if the underlying POD affects the non-conservative bias found in field estimation of POD from “hit/miss” type data. The first inspection technique represents a very good technique in terms of high SNR at the level of the detection threshold. In terms of the POD relationship, this results in a very steep curve. The second inspection technique represents a method with lower SNR at the level of the detection threshold, more representative of highly manual and operator-dependent NDI methods.

The simulation for POD estimation is the same as for the “hit/miss” data, except that inspections started at 1000 cycles.

I.3.2.1 “ \hat{a} vs. a ” Inspection Technique 1 – High SNR

The evolution of the in-service inspection data and estimated POD are shown in Figure I-12 through Figure I-14. It can be seen that at the earlier inspection time, which was the first inspection with any hits, the estimated POD from the field data is slightly non-conservative. Because of the steepness of the underlying POD, all the components are retired in the two next inspections. By the time of the inspection in Figure I-13, the field estimate of the POD is very close to the underlying POD.

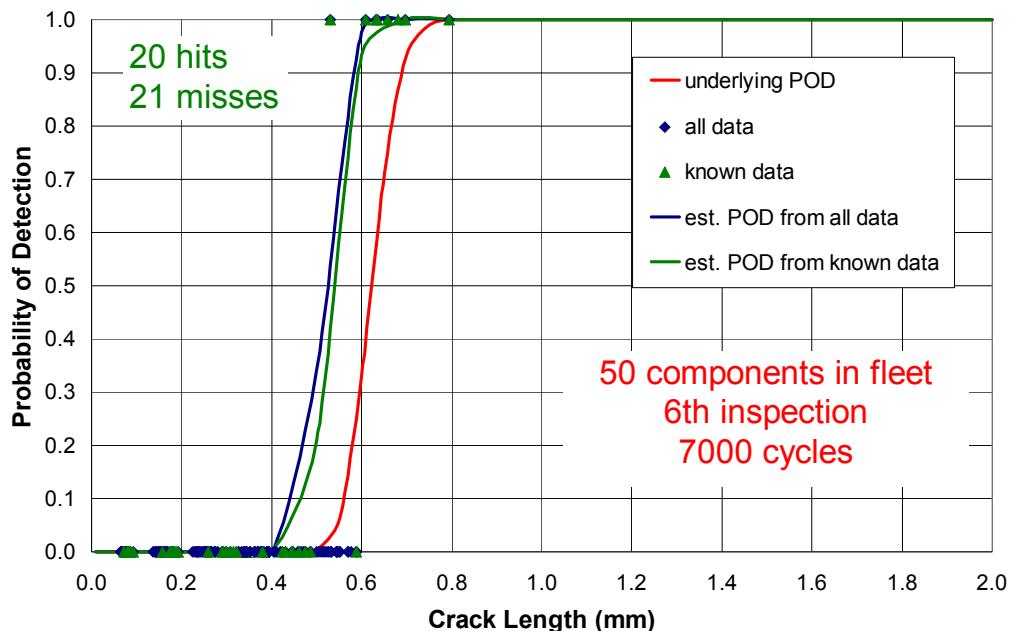


Figure I-12: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

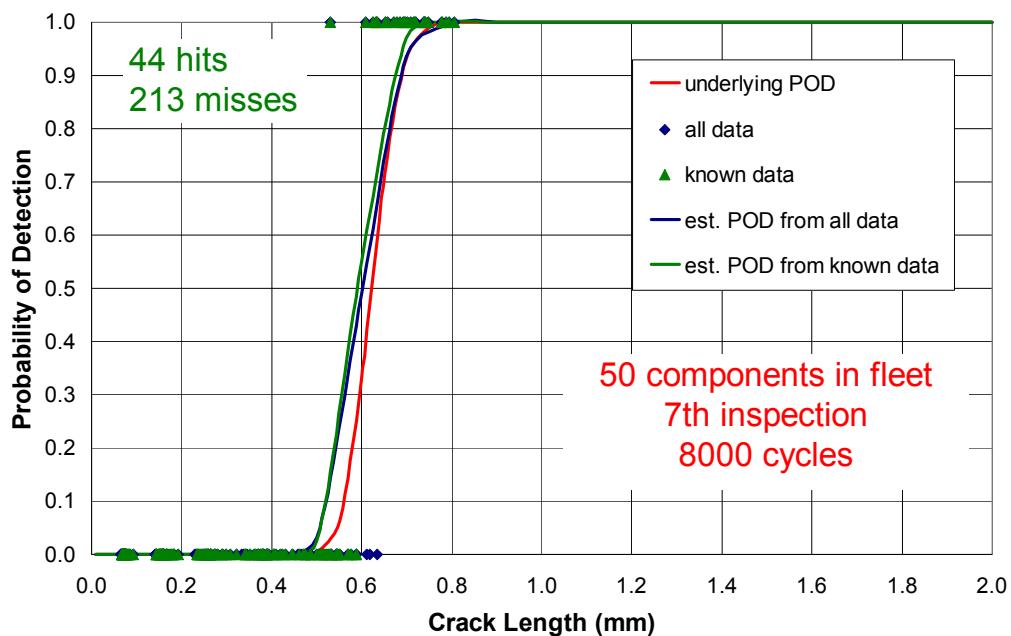


Figure I-13: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

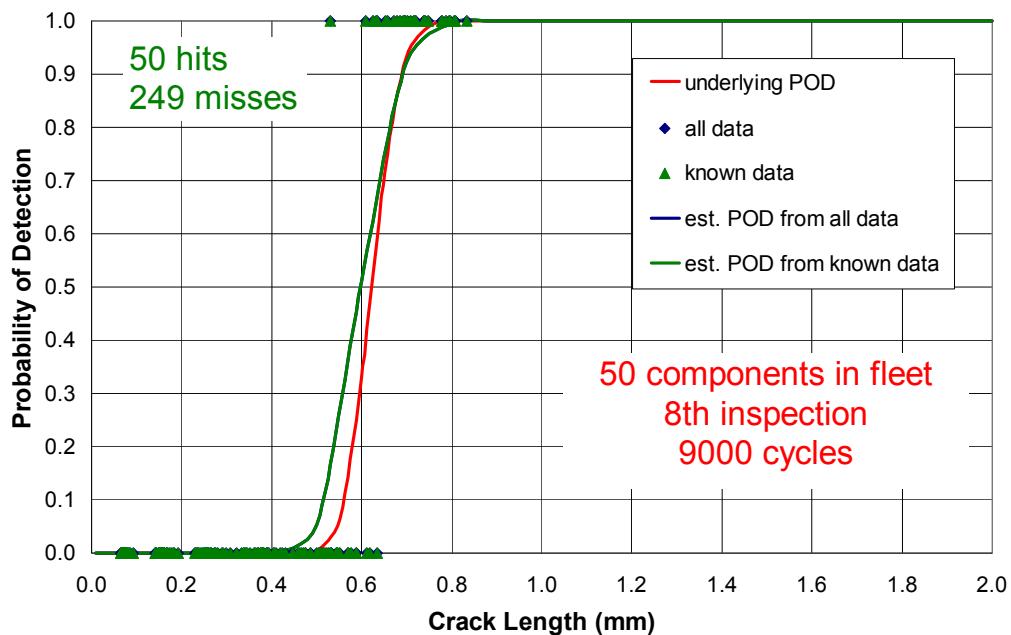


Figure I-14: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

I.3.2.2 “ \hat{a} vs. a ” Inspection Technique 2 – Low SNR

The evolution of the in-service inspection data and estimated POD are shown in Figure I-15 through to Figure I-18. At the earlier times, the POD estimate is very non-conservative. For this simulation, the estimated POD is still non-conservative when all the components in the fleet have been retired, as shown in Figure I-18 below.

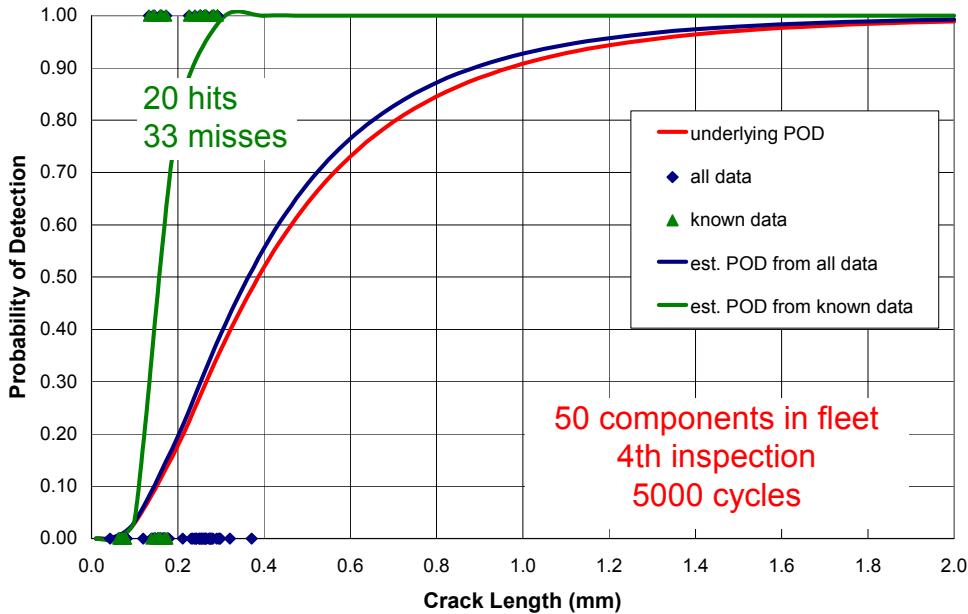


Figure I-15: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

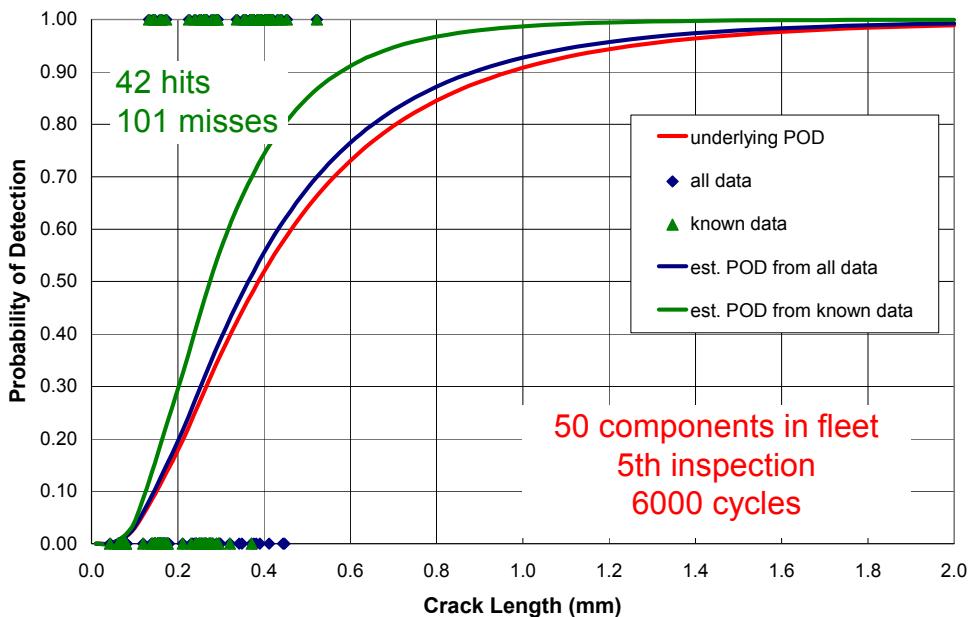


Figure I-16: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

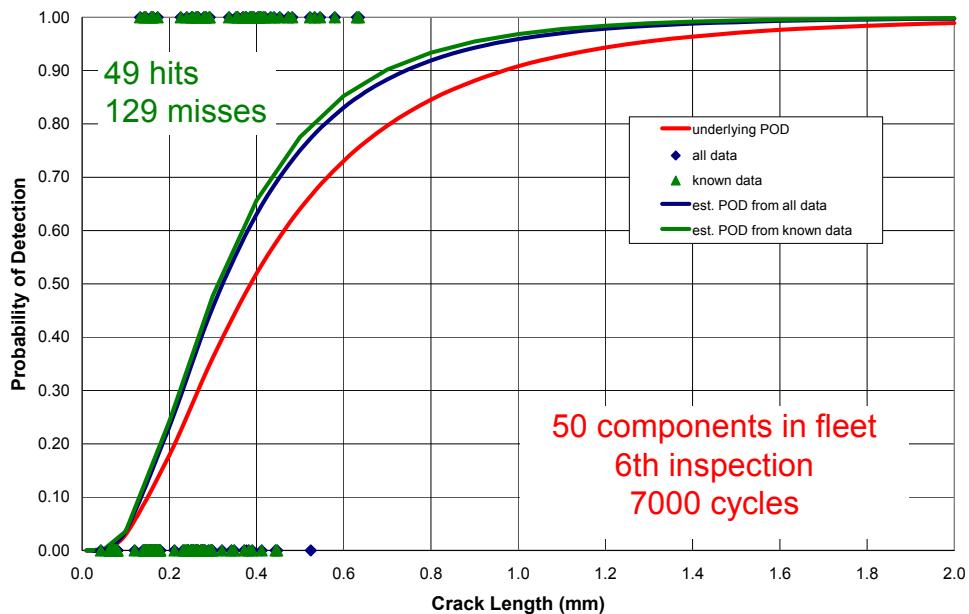


Figure I-17: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

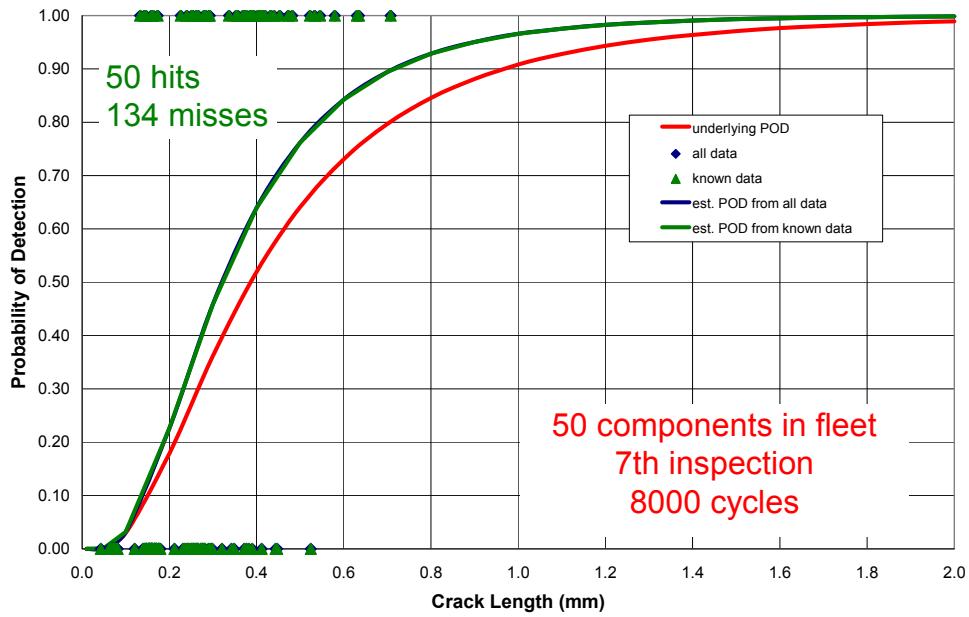


Figure I-18: “ \hat{a} vs. a ” Data and Estimated Mean PODs Compared with the Underlying POD.

The reason that the “ \hat{a} vs. a ” data gave a biased estimate for POD can be seen by examining Figure I-19. The cracks that are found first are those for which there is an unusually high signal, shown in green on Figure I-19. The POD curve is then estimated from a sub-population which has a larger response than the mean signal response, yielding a non-conservative estimate of the underlying POD.

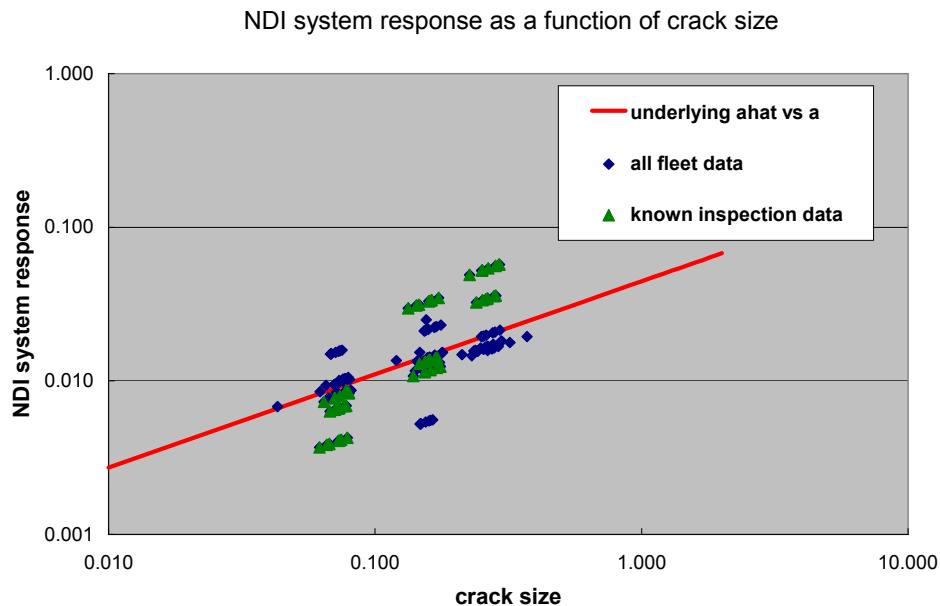


Figure I-19: The NDI System Response as a Function of Crack Size, for the Known Data from Inspection Findings, for All the Cracks in the Fleet, and the Underlying “ \hat{a} vs. a ” Relationship.

I.4 DISCUSSION

The simulations of both “hit/miss” and “ \hat{a} vs. a ”-type data showed that for high SNR inspection systems with steep POD curves, the estimates of POD from the simulated field data are very close to the underlying POD. However, lower SNR inspection systems with less steep POD curves yield very poor and non-conservative estimates of POD from field data. Unfortunately, highly manual techniques typical of field application tend to have relatively low SNR; and at this time, there is no way to determine a priori whether the field POD estimate is actually close to the underlying POD.

Crack populations, crack growth, inspection scheduling and the steepness of the underlying POD curve all affect the degree of non-conservatism found when attempting to estimate POD from field inspection data. In essence, in situations where inspections are repeated over time, the first cracks to be found are small cracks of low POD, as the crack size population moves towards larger crack sizes. Unless the inspection interval is very large, or the underlying POD is nearly vertical, most cracks will be found at sizes where POD is still low because of the multiple inspection opportunities at each site. Therefore, at any one time, the largest cracks in the field inspection data set are still going to be small compared to crack sizes at $POD \gg 0.5$. The POD curve fitting then will force the estimated POD to asymptotically approach unity at crack sizes for which the underlying POD is actually very small.

I.5 CONCLUSIONS

The motivation for investigation of POD estimation from field data was to accurately represent the human factors which can greatly affect POD. However, it has been shown theoretically by Spencer [7], and demonstrated experimentally in this paper, that resulting POD estimates are often non-conservative. The degree of non-conservatism is affected by crack size distributions, inspection intervals and the steepness

of the underlying POD, and without accurate knowledge of the underlying POD or of the actual crack size distribution it is impossible to determine the bias of a field data-based POD estimation.

As mentioned above, if the crack size distributions were known at the inspection times, then the POD could be corrected if necessary for bias. It is not known if it will be practically possible to determine crack size distributions accurately enough to do this. Some authors have also proposed Bayesian-based methods to try and overcome these difficulties.

I.6 REFERENCES

- [1] Simpson, D.L., *Development of Nondestructive Inspection Probability of Detection Curves Using Field Data*, Institute for Aerospace Research Report NRC-NAE-LTR-ST 1285, August, 1981.
- [2] Leemans, D.V., “Probability of Detection Based on Field Inspection Data”, Final Project Report for Department of National Defence (Canada) Air Vehicles Research Section, (1998).
- [3] Bruce, D.A., “NDT Reliability Estimation from Small Samples and In-Service Experience”, in *RTO Meetings Proceedings 10: Airframe Inspection Reliability under Field/Depot Conditions* (1998), RTO-MP-10, AC/323(AVT)TP/2, pp. 3-1 – 3-22.
- [4] Forsyth, D.S., Leemans, D., Fahr, A. and McRae, K., “Development of POD from In-Service NDI Data”, in the Review of Progress in QNDE, Vol. 19B, 2000, pp. 2167-2174.
- [5] Heida, J.H. and Grooteman, F.P., “Airframe Inspection Reliability using Field Inspection Data”, in *RTO Meetings Proceedings 10: Airframe Inspection Reliability under Field/Depot Conditions* (1998), RTO-MP-10, AC/323(AVT)TP/2, pp. 5-1 – 5-5.
- [6] Flood, J.T., Pfau, R.E., Grauvogl, E. and Regler, F., “Field Inspection Results and Damage Analysis of F4-F Horizontal Stabilizer Internal Structure”, in *RTO Meetings Proceedings 10: Airframe Inspection Reliability under Field/Depot Conditions* (1998), RTO-MP-10, AC/323(AVT)TP/2, pp. 16-1 – 16-8.
- [7] Spencer, F.W., Presentation to RTO-AVT-051 Committee Meeting, April, 2001.
- [8] Goranson, U.G., “Damage Tolerance: Facts and Fiction”, 14th Plantema Memorial Lecture, Presented at the 17th Symposium of the International Committee on Aeronautical Fatigue, Stockholm, Sweden, June 09, 1993.
- [9] Berens, A.P., “NDE Reliability Data Analysis”, in *Metals Handbook Volume 17: Nondestructive Evaluation and Quality Control* (9th ed.), ASM International, 1988.
- [10] Yee, B.G.W., Chang, F.H., Couchman, J.C., Lemon, G.H. and Packman, P.F., “Assessment of NDE Reliability Data”, NASA-CR-134991 (1975).
- [11] Berens, A.P. and Hovey, P.W., “Evaluation of NDE Reliability Characterization”, USAF Report No. AFWAL-TR-81-4160 (1981).
- [12] Spencer, F.W., “Identifying Sources of Variation for Reliability Analysis of Field Inspections”, in *RTO Meetings Proceedings 10: Airframe Inspection Reliability under Field/Depot Conditions* (1998), RTO-MP-10, AC/323(AVT)TP/2, pp. 11-1 – 11-8.

ANNEX I – SIMULATIONS OF THE EFFECT OF MISSING MISSES IN POD ESTIMATION FROM IN-SERVICE DATA

- [13] Koul, A.K., Fahr, A., Gould, G. and Bellinger, N.C., “Importance of Sensitivity and Reliability of NDI Techniques on Damage Tolerance Based Life Prediction of Turbine Discs”, in *AGARD/SMP Review: Damage Tolerance for Engine Structures 1. Non-Destructive Evaluation*, AGARD Report AGARD-R-768, 1988.
- [14] Kaprzynski, J.J., “Determination of Stress Intensity Factors in a J85 Can40 Fifth Stage Compressor Disc by Finite Element Analysis”, National Research Council Canada Report.

Annex J – AN APPROACH TO ESTIMATING POD CAPABILITIES FROM MAINTENANCE INSPECTION PROCEDURES

J.1 OVERVIEW

Maintenance data collected in past operations may not contain sufficient information to support development of POD capabilities. This method is applicable only to those methods that provide a scalar output such as ultrasonic and eddy current inspection methods. In most maintenance operations, actual flaw sizes detected is often missing or is estimated with considerable margin for error in flaw sizing. The performance capabilities may be estimated by development and transfer of signal and noise responses from known and representative artifacts if basis non-destructive inspection (NDI) parameters and reference artifacts are known and the NDI procedure can be reproduced with reasonable fidelity. Requirements for reproduction include:

- the inspection procedure including a documented acceptance criteria,
- the calibration artifacts (or duplicates),
- actual representative cracks of a size range that bound the expected detection threshold, and
- representative inspection equipment.

In short, it is necessary to reproduce the same data that is typically required to validate the applicability of an NDI procedure.

J.2 DATA COLLECTION AND OUTPUT

- A) Duplicate the equipment used in the inspection procedures.
- B) Duplicate the “calibration” – calibration must include at least three points.
- C) Make repetitive measurements on known cracks and background measurements in a location away from the cracks. Record scalar output (at least 29 measurements to provide a minimum confidence in bounding the measurement distribution – Figure J-1).
- D) Repeat the “calibration” using the selected artifacts (at least 29 measurements to provide a minimum confidence in bounding the measurement distribution).
- E) Plot the distribution and responses in the “calibration data” – this establishes a variance for the “calibration” process.
- F) Calculate an offset (transfer coefficient) from the “calibration values” for the crack measurements in the desired inspection geometry.
- G) Use the offset to plot a relationship between adjusted signal level and signal output (this is a most often a log-normal relationship).
- H) Apply the procedure specified ACCEPTANCE LEVEL to the resultant curve (Figure J-2).

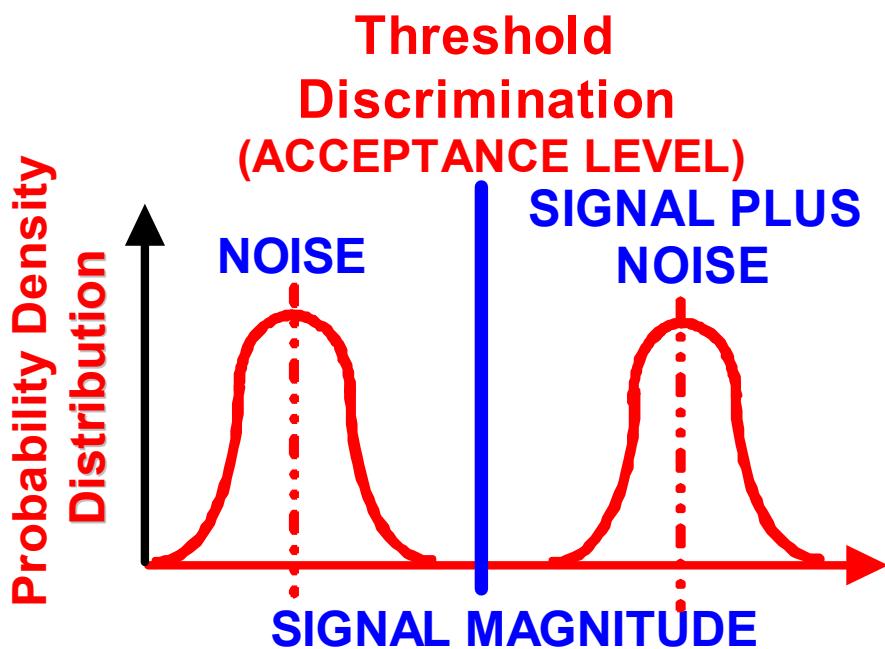


Figure J-1: Repetitive Signal and Noise Responses from Cracks of Equal Size.

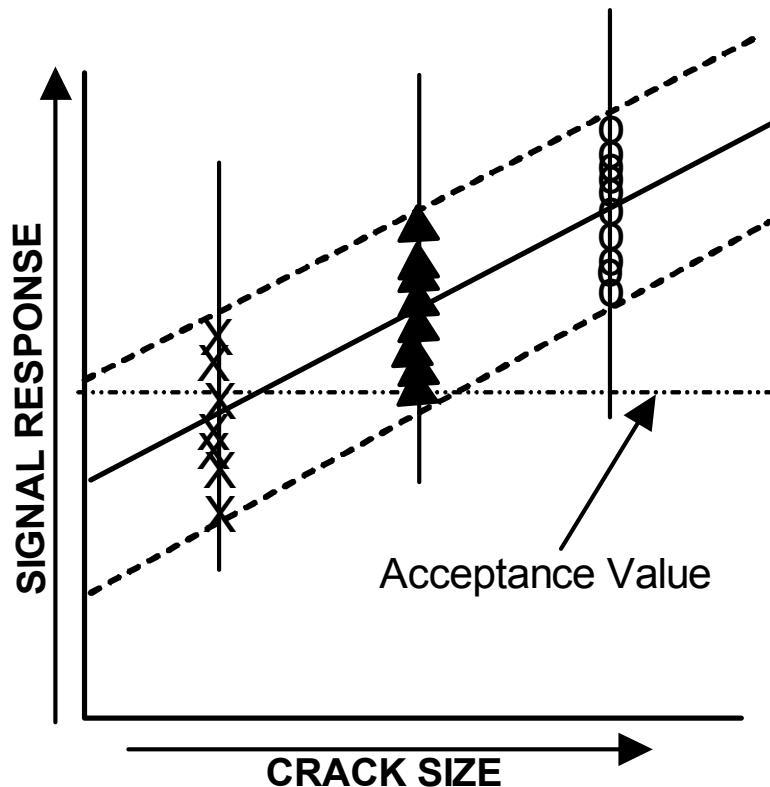


Figure J-2: Adjusted Response for Different Crack Sizes.

J.2.1 Plot Estimated POD Curves from Data Generated

The measured and bounded values shown in Figure J-2 provide the basis for generation of a POD output. Estimated POD curves may be generated by either the “hit/miss” or “ \hat{a} vs. a ” methods using a bounded offset from the data generated. The effects of adjusting the acceptance criteria can be readily observed by this method.

This method is often used to set acceptance criteria prior to **validation** of a specific inspection procedure. As a rule of thumb, the acceptance level should be set at that flaw size that provides a minimum of a three-to-one signal-to-noise response.

The method is only applicable to acceptance levels that are bounded by the “calibration” artifacts sizes and crack sizes within the bounded sizes. In addition, the method may not be applicable to small or very large crack sizes where the size of the probe/transducer is large with respect to the crack size interrogated. The method is only as good as the control of cracks, “calibration” artifacts and inspection procedures. Variance in any one of these factors will impact the validity and applicability of the method.

CAUTION: This method does not provide a validation POD demonstration and should be used *only* as an estimate of POD capability in the absence of adequate data. The method does not account for crack-to-crack response variances or *False Calls* due to human factors variables. The method is only applicable to those inspection modes that provide a scalar output for acceptance.

J.3 SUMMARY

POD generation from maintenance data requires:

- Precision in measurement of actual flaw sizes detected,
- Precision and rigid control of “calibration” artefacts,
- Precision and control of inspection procedures.

Much maintenance data does not contain information on the actual crack sizes detected, but simply rejects when the response exceeds a set acceptance level. When POD capabilities are desired from maintenance data, precision measurement and recording of detected crack sizes is required. In the absence of crack size measurement information, an alternate method is described to **estimate** POD capability using the “calibration” artifacts, representative cracks and specific inspection procedures as the basis for additional measurements and analyses. Fidelity of the method depends on cracks and “calibration” artifacts that are representative of the population, rigid control of “calibration” and the inspection procedure used for measurement.

Use of this procedure produces an **estimated** POD capability and should be used ***only*** as an estimate when more rigorous methods and/or data are not available.



REPORT DOCUMENTATION PAGE			
1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-TR-AVT-051 AC/323(AVT-051)TP/43	ISBN 92-837-1134-3	UNCLASSIFIED/ UNLIMITED
5. Originator	Research and Technology Organisation North Atlantic Treaty Organisation BP 25, F-92201 Neuilly-sur-Seine Cedex, France		
6. Title	The Use of In-Service Inspection Data in the Performance Measurement of Non-Destructive Inspections		
7. Presented at/Sponsored by	Work performed by the RTO Applied Vehicle Technology Panel (AVT) TG-051.		
8. Author(s)/Editor(s)	Multiple		9. Date March 2005
10. Author's/Editor's Address	Multiple		11. Pages 206
12. Distribution Statement	There are no restrictions on the distribution of this document. Information about the availability of this and other RTO unclassified publications is given on the back cover.		
13. Keywords/Descriptors	Aircraft Aircraft maintenance Airframes Damage assessment Data collection processes Data quality	Fatigue (materials) Inspection Laboratory NDI Life (durability) Life cycles Maintainability	Non-destructive tests POB (Probability of Detection) Reliability Service life Standards Verification inspection
14. Abstract	<p>Most available non-destructive inspection (NDI) reliability data results from analyses under laboratory or simulated in-service conditions. Adding analyses of in-service NDI findings could improve NDI. There is no organized process whereby in service data are collated for NDI reliability studies. The extent of this data and its usefulness to the NDI reliability program was analysed and processes for collection were defined as well as analytical methods to calculate NDI reliability from in-service data.</p> <p>Findings: One major contribution is a detailed summary of the close relationship between NDI, fracture mechanics and airworthiness including an important review of the statistical basis for many of current approaches to inspection.</p> <p>NDI maintenance records were reviewed for their usability to pooling of data from different sources in order to obtain statistically significant numbers. It was concluded that such records vary considerably in quality and fidelity and specific recommendations for improvement were made. Of the three approaches for using in-service inspection data to characterize the capability of an inspection system two aiming at characterizing NDI capability in terms of the probability of detection were rejected (POD) because in maintenance too many cracks of a detectable size are not detected rendering the basis data unreliable. A third approach summing inspection results in terms of the cumulative distribution function (CDF) of the sizes of the detected cracks does provide information about the capability of the NDI system in the in-service environment. The CDF does not directly yield the reliably detectable crack size (at a given confidence level) but it gives a first estimate of this size.</p>		





BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int



DIFFUSION DES PUBLICATIONS RTO NON CLASSIFIEES

Les publications de l'AGARD et de la RTO peuvent parfois être obtenues auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la RTO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus soit à titre personnel, soit au nom de votre organisation, sur la liste d'envoi.

Les publications de la RTO et de l'AGARD sont également en vente auprès des agences de vente indiquées ci-dessous.

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivi du numéro de série. Des informations analogues, telles que le titre est la date de publication sont souhaitables.

Si vous souhaitez recevoir une notification électronique de la disponibilité des rapports de la RTO au fur et à mesure de leur publication, vous pouvez consulter notre site Web (www.rta.nato.int) et vous abonner à ce service.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE

Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr (FIZBw)
Friedrich-Ebert-Allee 34, D-53113 Bonn

BELGIQUE

Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evêre, B-1140 Bruxelles

CANADA

DSIGRD2
Bibliothécaire des ressources du savoir
R et D pour la défense Canada
Ministère de la Défense nationale
305, rue Rideau, 9^e étage
Ottawa, Ontario K1A 0K2

DANEMARK

Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE

SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

ETATS-UNIS

NASA Center for AeroSpace
Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320

NASA Center for AeroSpace Information (CASI)

Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320
ETATS-UNIS

FRANCE

O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GRECE (Correspondant)

Defence Industry & Research
General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HONGRIE

Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ISLANDE

Director of Aviation
c/o Flugrad
Reykjavik

ITALIE

Centro di Documentazione
Tecnico-Scientifica della Difesa
Via XX Settembre 123
00187 Roma

LUXEMBOURG

Voir Belgique

NORVEGE

Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

AGENCES DE VENTE

The British Library Document Supply Centre

Boston Spa, Wetherby
West Yorkshire LS23 7BQ
ROYAUME-UNI

PAYS-BAS

Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE

Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL

Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE

LOM PRAHA s.p.
VTÚL a PVO o.z.
DIS ČR – NATO RTO
Tiskařská 8
100 38 Praha 10

ROYAUME-UNI

Dstl Knowledge Services
Information Centre, Building 247
Dstl Porton Down
Salisbury
Wiltshire SP4 0JQ

TURQUIE

Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanlıklar – Ankara

Les demandes de documents RTO ou AGARD doivent comporter la dénomination « RTO » ou « AGARD » selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications RTO et AGARD figurent dans les journaux suivants :

Scientific and Technical Aerospace Reports (STAR)

STAR peut être consulté en ligne au localisateur de ressources uniformes (URL) suivant:

<http://www.sti.nasa.gov/Pubs/star/Star.html>

STAR est édité par CASI dans le cadre du programme NASA d'information scientifique et technique (STI)
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
ETATS-UNIS

Government Reports Announcements & Index (GRA&I)

publié par le National Technical Information Service
Springfield

Virginia 2216

ETATS-UNIS

(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)



BP 25
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int



DISTRIBUTION OF UNCLASSIFIED RTO PUBLICATIONS

AGARD & RTO publications are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO reports, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your Organisation) in their distribution.

RTO and AGARD reports may also be purchased from the Sales Agencies listed below.

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number. Collateral information such as title and publication date is desirable.

If you wish to receive electronic notification of RTO reports as they are published, please visit our website (www.rta.nato.int) from where you can register for this service.

NATIONAL DISTRIBUTION CENTRES

BELGIUM

Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evêre
B-1140 Bruxelles

CANADA

DRDKIM2
Knowledge Resources Librarian
Defence R&D Canada
Department of National Defence
305 Rideau Street
9th Floor
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC

LOM PRAHA s.p.
VTÚL a PVO o.z.
DIS ČR – NATO RTO
Tiskárská 8
100 38 Praha 10

DENMARK

Danish Defence Research
Establishment
Ryvangs Allé 1
P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE

O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72
92322 Châtillon Cedex

GERMANY

Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)

Defence Industry & Research
General Directorate, Research Directorate
Fakinos Base Camp, S.T.G. 1020
Holargos, Athens

HUNGARY

Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND

Director of Aviation
c/o Flugrad, Reykjavik

ITALY

Centro di Documentazione
Tecnico-Scientifica della Difesa
Via XX Settembre 123
00187 Roma

LUXEMBOURG

See Belgium

NETHERLANDS

Royal Netherlands Military
Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY

Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND

Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL

Estado Maior da Força Aérea
SDFA – Centro de Documentação
Alfragide, P-2720 Amadora

SPAIN

SDG TECEN / DGAM
C/ Arturo Soria 289
Madrid 28033

TURKEY

Milli Savunma Bakanlığı (MSB)
ARGE ve Teknoloji Dairesi Başkanlığı
06650 Bakanlıklar – Ankara

UNITED KINGDOM

Dstl Knowledge Services
Information Centre, Building 247
Dstl Porton Down
Salisbury, Wiltshire SP4 0JQ

UNITED STATES

NASA Center for AeroSpace
Information (CASI)
Parkway Center, 7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

The British Library Document Supply Centre

Boston Spa, Wetherby
West Yorkshire LS23 7BQ
UNITED KINGDOM

Canada Institute for Scientific and Technical Information (CISTI)

National Research Council
Acquisitions
Montreal Road, Building M-55
Ottawa K1A 0S2, CANADA

NASA Center for AeroSpace

Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
UNITED STATES

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)

STAR is available on-line at the following uniform
resource locator:

<http://www.sti.nasa.gov/Pubs/star/Star.html>

STAR is published by CASI for the NASA Scientific
and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
UNITED STATES

Government Reports Announcements & Index (GRA&I)

published by the National Technical Information Service
Springfield

Virginia 2216

UNITED STATES

(also available online in the NTIS Bibliographic
Database or on CD-ROM)